



THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :
Laure TONINI

Le 2 décembre 2014

Titre :

Développement de nouvelles stratégies protéomiques pour l'analyse de
glycoprotéines mycobacteriennes

ED BSB : Biologie Structurale et fonctionnelle

Unité de recherche :

Institut de Pharmacologie et de Biologie Structurale – UMR 5089

Directeur(s) de Thèse :

Dr Odile SCHILTZ, Directrice de recherche au CNRS, Toulouse

Dr Michel RIVIERE, Directeur de recherche au CNRS, Toulouse

Rapporteurs :

Dr Sarah CIANFERANI - SANGIER, Chargée de recherche au CNRS, Strasbourg

Pr Jean-Louis HERRMANN, Professeur à l'Université de Versailles

Autre(s) membre(s) du jury :

Pr Pascal COSETTE, Professeur à l'Université de Rouen - Examineur

Pr Maryelle TROPIS, Professeur à l'Université de Toulouse - Présidente du jury

REMERCIEMENTS

Je souhaite tout d'abord remercier Odile et Michel pour leur encadrement au cours des trois dernières années.

Je remercie ensuite toutes les personnes que j'ai rencontrées au sein des deux équipes, que ce soit les personnes qui ont été présentes au cours des 3 ans ou celles qui l'ont été sur une plus courte durée et avec qui j'ai pu tout de même passer de très bons moments...

Je garderai d'ailleurs un très bon souvenir des pauses déjeuner, café ou goûter passées à discuter avec des personnes agréables. A ce titre je remercie alors Alexiane, Landry, Emeline, Mouin, Thomas M., Fanny, Luc, Alain, Martine, Jérôme et bien d'autres...

J'ai également passé 3 années en très bonne compagnie dans mon bureau, avec des personnes agréables et adorables que sont Carole, Carine et Nicolas. Votre présence quotidienne va à coup sûr me manquer !

Je remercie également Thomas L. pour sa disponibilité, sa gentillesse, son amitié au sein et en dehors du labo. Ta présence a beaucoup compté pour moi au cours de cette thèse.

Soso est incontestablement la personne du labo qui a été la plus importante pour moi au cours de cette thèse... Je la remercie grandement pour sa gentillesse, son optimisme et sa bonne humeur, son aide « pratique », et son amitié qui va également bien au-delà du labo.

Enfin, je remercie très chaleureusement mes parents. Ils ont toujours été très présents au cours de mes études et ont su me donner le soutien et la confiance en moi nécessaire pour aller jusque là...

Je remercie bien évidemment Ludo qui a été d'un très précieux soutien pour les bons et les mauvais moments de cette thèse...

Table des matières

Liste des principales abréviations.....	11
INTRODUCTION	13
Introduction première partie: La protéomique pour l'identification et la caractérisation de protéines, et cas particulier des protéines glycosylées.....	13
A. La préparation des échantillons pour une analyse de type nanoLC-MS/MS.....	14
I. Cas général.....	14
II. Cas des glycoprotéines : les méthodes d'enrichissement	16
1. La chromatographie d'affinité.....	17
2. La chromatographie d'interaction hydrophile	19
B. L'analyse nanoLC-MS/MS des peptides.....	20
I. La chromatographie liquide.....	20
II. L'analyse des peptides par spectrométrie de masse.....	21
1. Les principes généraux de la spectrométrie de masse appliquée à l'analyse de protéines.....	22
2. L'analyse de type « bottum-up ».....	24
3. L'appareillage utilisé.....	26
4. Cas particulier des peptides glycosylés	29
C. Le retraitement bio-informatique des données	33
I. Cas général.....	33
II. Cas des protéines glycosylées.....	35
1. Méthodes de glycoprotéomiques nécessitant une analyse préliminaires des protéines de l'échantillon	36
2. Méthodes de glycoprotéomiques ne nécessitant pas d'analyse préliminaires des protéines de l'échantillon	37
D. Les méthodes d'analyse de protéines entières pour la caractérisation des modifications post-traductionnelles	38
I. L'analyse de protéines entières par spectrométrie de masse pour l'identification des modifications.....	39
1. Fractionnement des protéines	39
2. L'introduction des protéines dans le spectromètre de masse.....	40

II. La stratégie d'analyse de type « top-down » pour la localisation des modifications post-traductionnelles	41
1. La fragmentation des protéines	42
2. Le retraitement bio-informatique des données.....	43
Introduction deuxième partie: L'application biologique.....	45
A. La tuberculose humaine	45
I. L'état actuel de la tuberculose dans le monde.....	45
II. Moyens de lutte	47
1. Le vaccin BCG.....	47
2. L'antibiothérapie	48
3. Souches MDR et XDR, et nouvelles drogues	50
B. <i>Mycobacterium tuberculosis</i> et sa pathogénicité	51
I. Structure de la bactérie	51
1. La membrane plasmique	51
2. La paroi et le complexe mAGP	52
3. La mycomembrane	52
4. La capsule	53
II. Physiopathologie des infections à <i>Mycobacterium tuberculosis</i>	53
III. Moyens de reconnaissance de <i>Mycobacterium tuberculosis</i> par le système immunitaire inné.....	55
1. Les récepteurs Toll-like.....	56
2. Les lectines de type C	56
C. La O-mannosylation des protéines mycobactériennes.....	58
I. Le mécanisme de O-glycosylation	58
II. Le rôle de la glycosylation des protéines chez <i>Mycobacterium tuberculosis</i>	60
PRESENTATION GENERALE DES TRAVAUX	63
RESULTATS Première partie : Analyse du glycoprotéome de <i>Mycobacterium tuberculosis</i> ..	65
A. Le contexte biologique	65
I. Impact de la glycosylation des protéines sur la viabilité et la croissance des souches Δ Msmeg_5447 et Δ Rv1002c.....	65
II. Impact de la mannosylation des protéines sur la virulence de la souche Rv1002c	66
B. Identification des glycoprotéines sécrétées de <i>Mycobacterium tuberculosis</i> – Milieu de culture 7H9	68

I. Echantillons biologiques	68
II. Conditions analytiques	69
III. Evaluation de l'apport d'une méthode d'égalisation sur l'identification des protéines 71	
IV. Le retraitement bio-informatique des données pour l'identification de glycoprotéines.....	72
1. La recherche des protéines glycosylées avec le moteur de recherche Mascot	73
2. Principe du logiciel « Neutral Loss ».....	73
3. Paramétrisation du logiciel.....	74
4. Retraitement des données issues du logiciel « Neutral Loss »	75
V. Résultats	77
1. Apport de l'outil bioinformatique Neutral Loss pour la sélection des spectres de glycopeptides putatifs	77
2. Glycoprotéines identifiées.....	78
VI. Conclusion	81
C. Essais d'enrichissement	81
I. Enrichissement par affinité à la Concanavaline A.....	81
1. Enrichissement au niveau protéique.....	81
2. Enrichissement au niveau peptidique	83
II. Enrichissement par chromatographie d'interaction hydrophile	84
D. Recherche des glycoprotéines sécrétées de <i>Mycobacterium tuberculosis</i> – Milieu de culture Sauton	85
I. Echantillons biologiques et conditions analytiques.....	85
II. Retraitement des données	86
III. Résultats et conclusion	87
E. Recherche des glycoprotéines sécrétées de <i>Mycobacterium tuberculosis</i> – Milieu de culture Sauton, souches mutante et complémentée.....	92
I. Echantillons biologiques, conditions analytiques et retraitement des données.....	92
1. Analyse des protéines de la souche mutante	93
2. Analyse des protéines de la souche complémentée.....	93
F. Protéines cellulaires	97
G. Apport de l'outil « Neutral Loss » par rapport à une stratégie de recherche classique des glycopeptides avec le moteur de recherche Mascot	100
I. Confrontation de deux approches.....	100

II. Résultats	100
H. Conclusions & perspectives	102
RESULTATS deuxième partie: Recherche des voies métaboliques de <i>Mycobacterium tuberculosis</i> impactées par un arrêt du processus de glycosylation	105
A. Les méthodes de quantification relative	105
I. Les méthodes avec marquage isotopique	105
II. Les méthodes sans marquage isotopique ou « label-free »	106
1. Quantification sans marquage par extraction du signal MS	106
2. L'analyse bio-informatique des données de protéomique quantitative sans marquage par extraction du signal MS.....	108
3. Quantification sans marquage par « spectral counting »	108
4. L'analyse bio-informatique des données de protéomique quantitative sans marquage par extraction du signal MS.....	109
III. Les méthodes de quantification absolue.....	110
B. Résultats.....	110
I. Conditions analytiques	110
II. Protéines validées.....	111
III. Protéines différentiellement exprimées.....	112
1. Protéines du surnageant de culture	113
2. Protéines issues de la lyse des bactéries.....	115
IV. Conclusion	121
RESULTATS troisième partie: Mise en place de méthodes d'analyse par spectrométrie de masse pour la caractérisation de la glycosylation de la Fasciline de <i>Mycobacterium smegmatis</i>	123
A. Contexte.....	123
B. Détermination des modifications post-traductionnelles de la protéine	124
I. Mise en place d'une analyse de la protéine entière pour la détermination des protéoformes	124
II. Recherche des peptides modifiés via une approche de type « bottom-up ».....	126
C. Caractérisation des sites de glycosylation de la protéine.....	130
I. Mise en place d'une analyse de type « top-down » pour une caractérisation des fragments modifiés de la protéine.....	131
1. Conditions analytiques	131
2. Résultats – fragmentation SID.....	132

3. Résultats – fragmentation HCD	134
4. Résultats – fragmentation ETD.....	135
5. Conclusion	135
II. Recherche des sites de glycosylation via une approche de type « bottom-up » avec fragmentation ETD	135
1. Résultats – peptide N-terminal	135
2. Résultats – peptide C-terminal.....	137
III. Recherche des sites de glycosylation via une approche de type « bottom-up » sur des peptides aux motifs saccharidiques « simplifiés »	139
IV. Recherche des sites de glycosylation par la technique de spectrométrie de masse de mobilité ionique	141
1. Principe.....	141
2. Résultats	142
D. Conclusion	146
CONCLUSION GENERALE ET PERSPECTIVES	149
MATERIEL ET METHODES.....	155
I. Préparation des échantillons pour l'analyse par spectrométrie de masse: digestion en gel	154
II. Analyse nanoLC-MS/MS et retraitement des données de glycoprotéomique et de protéomique quantitative.....	156
III. Analyse ESI-MALDI des peptides	158
IV. Enrichissement en glycoprotéines ou glycopeptides	158
V. Analyse des protéines entières	159
PUBLICATION & COMMUNICATIONS	161
Liste des communications	161
Bibliographie.....	178

Liste des principales abréviations

BCG : Bacille de Calmette & Guérin

CID : « collision induced dissociation »

ConA : Concanavaline A

DO : Densité Optique

ESI : Electrospray ou électronébulisation

ETD : Electron transfer dissociation

HCD : Higher-energy Collisional

HILIC : Hydrophilic liquid chromatography

MALDI : Matrix-Assisted Laser Desorption/Ionisation

MS/MS : spectrométrie de masse en tandem

Mtb : *Mycobacterium tuberculosis*

m/z : ratio masse/charge

PFF : « Peptide Fragment Fingerprinting

PTM : Modification post-traductionnelle

SDS : dodécylsulfate de sodium

INTRODUCTION

Introduction première partie: La protéomique pour l'identification et la caractérisation de protéines, et cas particulier des protéines glycosylées

Les modifications post-traductionnelles (PTM) des protéines entraînent la modification de leur structure et/ou de leur fonction. Les PTMs sont donc importantes dans la définition du rôle biochimique des protéines dans certains processus cellulaires.



Figure 1: Environ 5 millions de protéines sont codées par les 20 000 gènes du génome humain

Près de 400 PTMs différentes sont à ce jour répertoriées (www.uniprot.org/docs.ptmlist). La caractérisation et la connaissance de ces différentes PTMs est essentielle pour l'étude des systèmes biologiques.

Afin par exemple de comprendre le rôle d'une protéine donnée dans un processus cellulaire, un des principaux défis est de parvenir à caractériser l'ensemble des PTMs de celle-ci, mais aussi de les localiser sur la séquence en acides aminés afin de comprendre le rôle de chacune d'entre elles dans la modulation de certains rôles fonctionnels de la protéine. La spectrométrie de masse est la technique la plus utilisée qui permette aujourd'hui la caractérisation et la localisation des PTMs.

Parmi la multitude de PTMs référencées à ce jour, la glycosylation des protéines est considérée comme la modification la plus abondante, avec des effets considérables sur les propriétés biologiques des protéines (Spiro (2002)). Cette PTM a longtemps été pensée comme restreinte aux eucaryotes (avec près de 50% des protéines portant cette modification), mais il a été démontré qu'elle est également retrouvée chez les procaryotes, notamment chez *Mycobacterium tuberculosis*, l'agent étiologique de la tuberculose humaine.

L'étude de protéines glycosylées par des approches de spectrométrie de masse et de protéomique (l'étude de l'ensemble des protéines présentes dans une cellule, un tissu, un organe ou un organisme à un moment précis et sous des conditions données, Kahn 1995) demeure un défi technologique (Pan, Chen et al. 2011). L'étude des protéines par une stratégie protéomique classique se compose de trois parties (figure 2): une première partie de préparation de l'échantillon (obtention d'un échantillon peptidique à partir d'un échantillon biologique), une deuxième partie d'analyse des peptides par nanoLC-MS/MS (fractionnement des peptides par chromatographie liquide et analyse des peptides par spectrométrie de masse), et une troisième partie d'analyse bioinformatique des données acquises pour l'identification, la caractérisation et/ou la quantification des protéines.

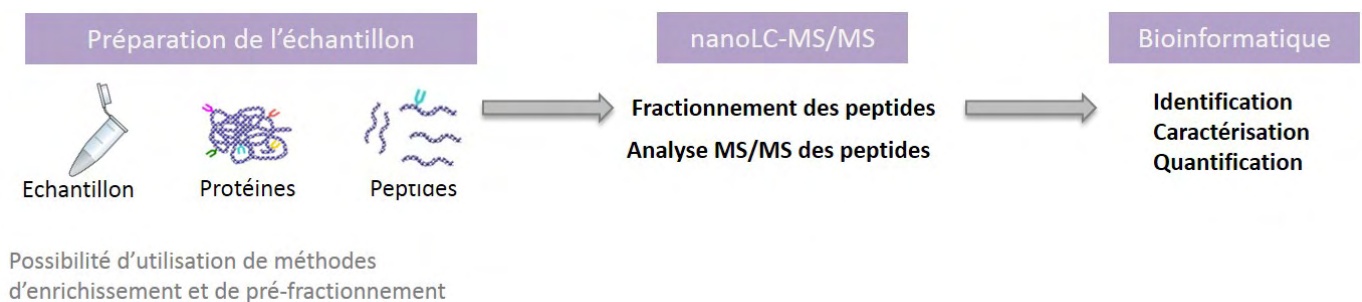


Figure 2: Les différentes étapes d'une analyse protéomique.

Or, pour l'étude des protéines glycosylées, des difficultés sont rencontrées dès l'étape de préparation des échantillons. Les protéines glycosylées possèdent des propriétés biochimiques différentes des protéines non modifiées, et peuvent également être présentes en faible abondance dans un échantillon d'intérêt. Il est alors nécessaire d'utiliser et de concevoir des stratégies de préparation d'échantillon adaptées pour étudier ce type de modification. De plus, la recherche à haut-débit de ces protéines glycosylées dans un échantillon biologique complexe par une stratégie protéomique classique souffre encore aujourd'hui d'un manque d'outils bioinformatiques dédiés et performants. Enfin, la détermination précise des acides aminés porteurs des modifications nécessite des technologies de spectrométrie de masse innovantes.

A. La préparation des échantillons pour une analyse de type nanoLC-MS/MS

I. Cas général

Au cours d'une étude de protéomique, les méthodes de préparation d'échantillon et de fractionnement des protéines à utiliser doivent être des méthodes permettant l'extraction de protéines aux propriétés biochimiques diverses afin de ne pas discriminer une sous-

population de protéines quelconque. D'autre part, les méthodes utilisées doivent également permettre l'analyse de protéines minoritaires pour une détermination du protéome en profondeur. Pour cela, il est possible d'utiliser plusieurs méthodes de fractionnement sur le même échantillon de départ. Dans ce but, la méthode « Mudpit » de Yates consiste par exemple en des chromatographies multidimensionnelles (Washburn, Wolters et al. 2001). Cependant, le nombre de fractions à analyser par la suite et donc le temps d'analyse est important. De plus, multiplier les étapes de préparation d'échantillon introduit des sources de variabilité dans le mode opératoire (Antberg, Cifani et al. 2012).

Un bon compromis, le plus couramment utilisé en protéomique, est d'utiliser une méthode de fractionnement des protéines sur gel d'électrophorèse à une dimension, puis un fractionnement des peptides par chromatographie liquide.

Le fractionnement sur gel d'électrophorèse à une dimension en conditions dénaturantes, ou gel SDS-Page est une des méthodes de fractionnement des protéines les plus répandues (figure 3). Elle est basée sur le déplacement des protéines chargées négativement (SDS fixé sur les protéines) au travers d'un gel de polyacrylamide, sous l'effet d'un champ électrique. En fonction du caractère physico-chimique de chacune des protéines, celles-ci acquièrent des vitesses de migration différentes et sont séparées les unes des autres. Cette technique découle des travaux de Laemmli (Laemmli 1970). Cette technique peut présenter des limites en termes de résolution pour des échantillons complexes. En effet, le fractionnement de protéines sur colonne échangeuse de cations a par exemple permis l'identification d'un plus grand nombre de protéines à partir d'une lysat de *Escherichia Coli* qu'avec un fractionnement sur gel 1D SDS-Page, dans les mêmes conditions analytiques (Mostovenko, Hassan et al. 2013).

En revanche, cette technique de fractionnement est appréciée pour la séparation des protéines difficiles à solubiliser (protéines membranaires par exemple) pour lesquelles il est difficile d'utiliser un gel bi-dimensionnel (nécessité d'utiliser des détergents doux qui ne permettent pas la solubilisation de ces protéines).

La digestion des protéines en peptides est ensuite réalisée *in gel*, le plus souvent en utilisant la trypsine dont le poids moléculaire lui permet de pénétrer les mailles du gel et digérer les protéines enchâssées dans la matrice d'acrylamide. C'est une protéase stable qui clive les protéines en peptides de manière très spécifique du côté C-terminal des résidus lysine et arginine (Olsen, Ong et al. 2004), ce qui facilite par la suite l'identification des peptides par recherche dans des bases de données. Les peptides sont ensuite eux-mêmes fractionnés par chromatographie liquide en amont de l'analyse par spectrométrie de masse.

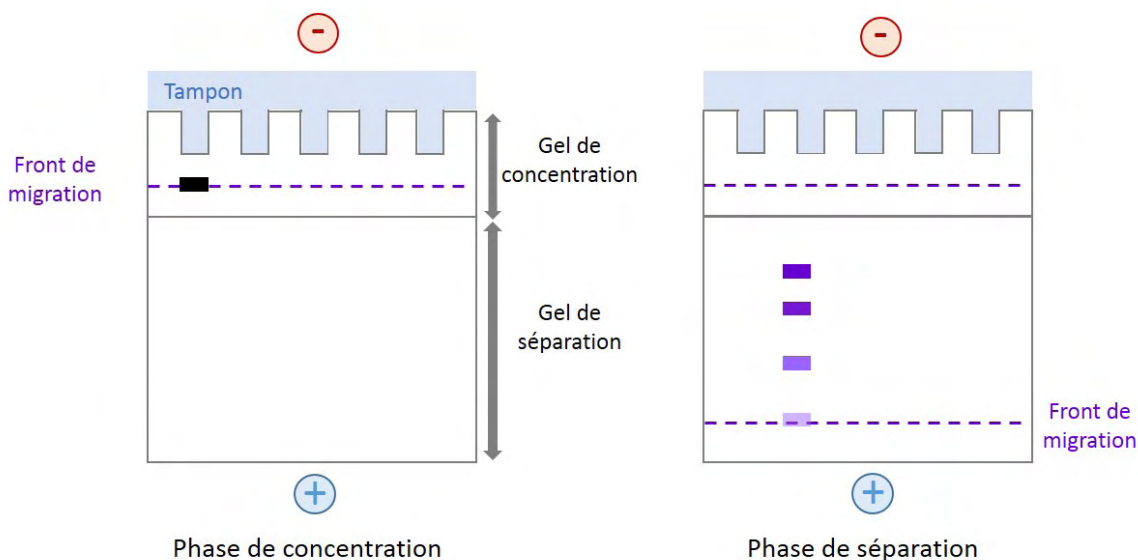


Figure 3: Le fractionnement des protéines sur un gel SDS-Page, avec les phases de concentration de séparation. Au sein du gel de concentration (avec un pourcentage d'acrylamide de 4%), les protéines sont concentrées mais ne sont pas séparées selon leurs poids moléculaire. Au sein du gel de séparation, les protéines sont séparées jusqu'au front de migration.

II. Cas des glycoprotéines : les méthodes d'enrichissement

L'analyse protéomique pour la recherche de protéines glycosylées profite du développement de méthodes d'enrichissement qui peuvent permettre de décomplexifier l'échantillon en amont de l'analyse des peptides par spectromètres de masse.

Dans cette thèse, je me suis plus particulièrement intéressée à la recherche et à la caractérisation de protéines *O*-glycosylées mycobactériennes, avec des mannoses comme motifs glycaniques.

Certaines des méthodes d'enrichissement disponibles à ce jour sont spécifiques d'un type de glycosylation ou peuvent être utilisées pour tous les types de glycosylation. Pour une analyse nanoLC-MS/MS des protéines glycoylées, plusieurs stratégies analytiques sont alors possibles. En effet, il est possible d'utiliser une technique d'enrichissement au niveau protéique ou au niveau peptidique, ou bien aux deux niveaux, ou encore d'analyser une fraction peptidique totale, non enrichie (figure 4).

Les méthodes d'enrichissement aujourd'hui les plus utilisées sont la chromatographie d'affinité (affinité des glycoprotéines ou glycopeptides pour la protéine concanavaline A, ou pour des phases fonctionnalisées avec certains groupements chimiques) ou la chromatographie d'interaction hydrophile.

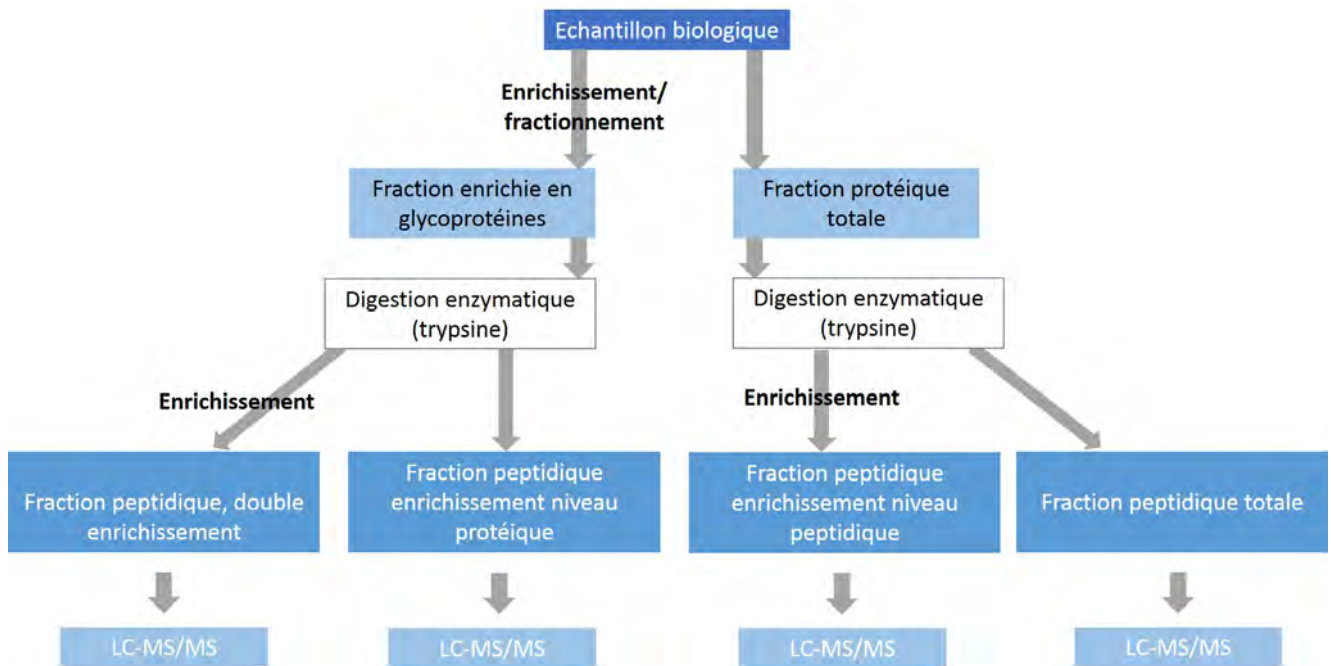


Figure 4: Les différentes stratégies possibles pour l'analyse de glycopeptides au sein de mélanges complexes. Il est possible d'analyser l'ensemble des peptides du mélange, ou d'utiliser des méthodes d'enrichissement au niveau protéique ou peptidique.

1. La chromatographie d'affinité

- **Les phases de lectines:**

L'enrichissement sur phases de lectines est basé sur la propriété de ces protéines à se lier de façon spécifique et réversible à certains motifs glycaniques.

Il existe de nombreuses lectines qui reconnaissent différents glycanes, dont les plus connues sont la concanavaleine A qui reconnaît spécifiquement les résidus mannosylés ou glucosylés (Becker, Reeke et al. 1975) ou la WGA (Wheat Germ Agglutinin) qui possède une affinité pour les acides sialiques et les motifs *N*-acetyl-glucosamine. Plusieurs lectines peuvent être combinées afin d'améliorer le rendement d'enrichissement ou augmenter la variété de glycoformes à enrichir.

Le plus souvent, les lectines sont fixées sur de la sepharose ou de l'agarose dans une colonne de chromatographie pour purifier les protéines (Espitia and Mancilla 1989) ou les peptides glycosylés. Dans le cas des *N*-glycosylations, les protéines potentiellement glycosylées ayant une affinité à la lectine sont « décrochées » de la colonne suite à l'ajout d'un compétiteur.

Cette technique peut être utilisée pour des protéines *O*-glycosylées. Cependant, du fait du caractère réversible de la liaison de la lectine aux motifs glycaniques, cette méthode d'enrichissement souffre d'un faible rendement de purification. De plus, des interactions non-spécifiques de composés non-glycosylés avec les lectines sont souvent observées.

- **Les phases hydrazides:**

L'enrichissement de glycopeptides ou glycoprotéines sur phase d'hydrazide se déroule en deux étapes : l'oxydation des diols vicinaux *cis* d'oligosaccharides en aldéhydes est d'abord effectuée avec le periodate de sodium, et les glycoprotéines ou glycopeptides sont ensuite immobilisés sur la phase via une réaction chimique donnant lieu à une liaison covalente entre les phases hydrazides et les aldéhydes (figure 5) (Zhang, Li et al. 2003). Pour les motifs *N*-glycanniques, une étape de déglycosylation (avec une enzyme spécifique du motif glycanique étudié ou avec la PNGase F, active sur tous les types de motifs, excepté chez les plantes) est ensuite nécessaire afin de libérer les glycopeptides ou glycoprotéines de la colonne. Sur les motifs *O*-glycanniques, une étape chimique de β -élimination peut être envisagée.

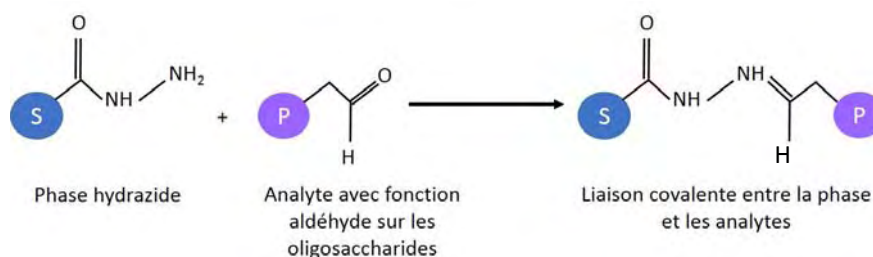


Figure 5: Réaction de type substitution nucléophile entre l'amine secondaire de la phase hydrazide et la fonction aldéhyde de l'analyte

Les groupements aldéhydes étant généralement absents des peptides ou protéines non glycosylés, cette technique est relativement spécifique. De plus, les liaisons entre les motifs glycaniques et la phase étant de nature covalente, les risques de décrocher les glycopeptides ou glycoprotéines au cours d'éventuelles étapes de lavage sont plus faibles. Ainsi, la multiplication de ces étapes peut permettre de réduire le taux de co-purification de peptides ou protéines non-glycosylés.

- **La phase d'acide boronique**

Une phase fonctionnalisée avec de l'acide boronique peut également être utilisée pour enrichir un échantillon en glycopeptides ou glycoprotéines. En effet, une liaison covalente se forme entre les fonctions *cis*-diols d'un motif glycosidique et l'acide boronique en milieu acide, et cette réaction peut être réversée en milieu basique (Rawns and Lienhard 1974).

- **La phase graphite**

Il est également possible d'utiliser une phase graphite pour l'enrichissement en peptides de petite taille et en peptides hydrophiles, donc en glycopeptides (Larsen, Cordwell et al. 2002, Xu, Wu et al. 2009). Cette technique est une variante de la chromatographie de phase inverse. Elle permet de retenir les peptides hydrophiles qui peuvent être perdus au cours d'une chromatographie de phase inverse classique.

2. La chromatographie d'interaction hydrophile

Les glycopeptides et glycoprotéines sont généralement plus polaires et hydrophiles que leurs correspondants non-glycosylés du fait des groupements hydroxyles qu'ils possèdent. Ce caractère peut être exploité au cours d'une chromatographie d'interaction hydrophile (HILIC, Hydrophilic liquid chromatography). La chromatographie d'interaction hydrophile est une variante de la chromatographie en phase normale pour laquelle la phase stationnaire est une phase hydrophile (figure 6) (Alpert 1990). Les analytes sont retenus sur la phase stationnaire via des liaisons hydrogènes, des interactions ioniques ou des interactions dipôle-dipôle.

Plusieurs phases peuvent être utilisées dans le cas de cette chromatographie, les phases ZIC-HILIC (Zwitterionic Hydrophilic interaction Liquid Chromatography) (Takegawa, Ito et al. 2008), ou des billes de silice ou sépharose modifiées (Wollscheid, Bausch-Fluck et al. 2009). Il est possible d'utiliser des colonnes chromatographiques HILIC mais aussi des micro-colonnes d'extraction remplies de phase HILIC permettant l'enrichissement d'un échantillon en peptides hydrophiles.

Plusieurs études ont montré que cette technique permet un enrichissement en peptides *N*-glycosylés, mais aussi *O*-glycosylés (Wuhrer, de Boer et al. 2009, Zauner, Koeleman et al. 2010).

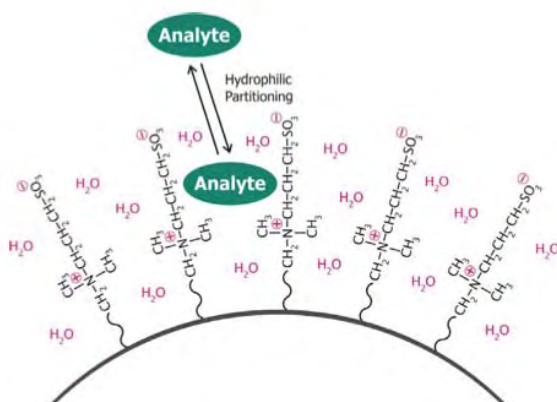


Figure 6: La phase zwitterionique ZIC-HILIC

En revanche, du fait de l'utilisation d'éluants avec des concentrations élevées en solvants organiques, un problème de solubilité des peptides peut être rencontré. De plus, cette technique d'enrichissement en peptides hydrophiles n'est pas spécifique des peptides glycosylés, des peptides hydrophiles non glycosylés peuvent alors être élués parmi les peptides glycosylés. En effet, le caractère hydrophile d'un peptide apporté par le motif glycosidique qu'il porte peut être plus faible que le caractère hydrophile apporté par les acides aminés d'autres peptides non-glycosylés, notamment dans le cas de peptides O-glycosylés possédant de courts motifs glycosidiques, tels que les peptides glycosylés qui seront étudiés au cours de cette thèse.

D'autres types de chromatographie ou de techniques d'enrichissement ont été cités dans la littérature. Un peptide glycosylé possédant une géométrie différente et plus encombrante que son homologue non-glycosylé, la chromatographie d'exclusion stérique peut par exemple être envisagée (Alvarez-Manilla, Atwood et al. 2006).

L'utilisation de méthodes d'enrichissement en amont de l'analyse des peptides par spectrométrie de masse peut alors permettre une analyse aisée des peptides glycosylés, dans le cas du choix d'une méthode adaptée au type de glycosylation étudié. Nous verrons dans la suite de ce travail l'évaluation de l'apport de deux des méthodes ici décrites pour l'étude de peptides O-glycosylés mycobactériens.

B. L'analyse nanoLC-MS/MS des peptides

I. La chromatographie liquide

La chromatographie liquide est une technique de séparation des analytes entre une phase liquide et une phase stationnaire. La phase mobile est constituée d'un mélange de solvants et la phase stationnaire peut être constituée d'une résine ou d'un gel.

La chromatographie d'exclusion stérique, la chromatographie d'échange d'ions, la chromatographie de phase inverse, la chromatographie d'interaction hydrophile et la chromatographie d'affinité constituent les techniques de chromatographie liquide qui peuvent être utilisées en protéomique et permettent de séparer les peptides et les protéines en fonction de différents paramètres physico-chimiques.

La chromatographie liquide est une technique séparative intéressante en raison de sa résolution, sa flexibilité et ses possibilités de couplage direct aux spectromètres de masse via la source ESI (dans le cas de l'utilisation de tampons compatibles avec une analyse par spectrométrie de masse).

La chromatographie de phase inverse permet la séparation des peptides en fonction de leur hydrophobicité et de leur taille. La phase stationnaire est une phase apolaire, souvent constituée de silices apolaires greffées, de deux à dix-huit carbones. La séparation des peptides est généralement effectuée grâce à un gradient d'élution constitué d'un mélange eau/acétonitrile/acide généralement. L'hydrophobicité de l'éluant varie de façon continue. Les peptides sont retenus dans la colonne en fonction de leur hydrophobicité, puis élués à un certain temps de rétention sous la forme d'un pic chromatographique.

Jusqu'au début des années 2000, des colonnes avec des particules entre 3 et 5 μm et une porosité de 300 Å étaient les plus utilisées. Depuis, des colonnes avec des particules plus petites sont apparues, de l'ordre du micron. Ces colonnes permettent des séparations plus efficaces avec de meilleures résolutions mais entraînent une augmentation de la pression au sein des colonnes, de l'ordre de 600-800 bars (Shen, Smith et al. 2005). Des systèmes adaptés ont alors été développés afin de supporter de telles pressions et compatibles avec les sources nano-ESI.

II. L'analyse des peptides par spectrométrie de masse

1. Les principes généraux de la spectrométrie de masse appliquée à l'analyse de protéines

La spectrométrie de masse est une méthode analytique physico-chimique et destructive permettant la mesure précise de la masse moléculaire d'un analyte, composé chimique ou biologique, en mesurant un rapport masse sur charge (m/z , en thomson). Cette technique est née des travaux de Joseph John Thomson (prix Nobel de physique en 1906) et de Francis William Aston (prix Nobel de chimie en 1922).

Un spectromètre de masse est classiquement représenté par trois éléments : une source d'ionisation, un analyseur de masse et un détecteur de courant d'ions (figure 7).

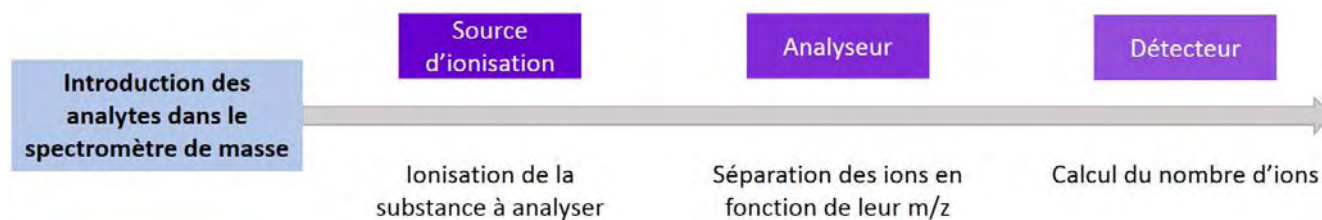


Figure 7: Les trois éléments composant un spectromètre de masse (source d'ionisation, analyseur et détecteur)

La source d'ionisation permet l'obtention des analytes sous forme d'ions en phase gazeuse. Les ions sont ensuite focalisés et transmis à l'analyseur dans lequel ils sont séparés en fonction de leur rapport m/z . En effet, deux ions de m/z différents vont avoir un comportement différent au sein d'un champ électrostatique ou électromagnétique (Domon and Aebersold 2006). Les ions sont ensuite collectés par le détecteur qui mesure leur nombre. A l'issue de la détection, un système informatique traite les données et permet l'obtention de l'information sous la forme de spectres de masse, avec l'intensité de chacun des ions en ordonnées et le rapport m/z de chacun d'entre eux en abscisses.

Les deux dernières étapes, la séparation des ions dans l'analyseur et la mesure du signal dans le détecteur se font dans un vide poussé ($10^{-4} - 10^{-7}$ mbar) pour que les ions puissent se déplacer librement sans entrer en collision avec des molécules de gaz et parvenir intacts au niveau du détecteur.

Les peptides et les protéines sont des composés thermiquement instables, non-volatiles et polaires. Deux sources d'ionisation permettent alors l'étude de macromolécules biologiques et n'ont été développées qu'au cours des années 80 : la source électrospray ou électronébulisation (ESI) et la source MALDI (*Matrix-Assisted Laser Desorption/Ionisation*). Les deux types d'ionisation permettent d'analyser des échantillons à des concentrations de l'ordre du sub-picomolaire.

L'ionisation ESI a été développée par le groupe de Fenn (prix Nobel de chimie en 2002, (Fenn, Mann et al. 1989)) et s'utilise pour l'analyse d'échantillons sous forme liquide à pression atmosphérique. Cette source d'ionisation peut donc être couplée à des techniques séparatives telles qu'un système de chromatographie liquide (Whitehouse, Dreyer et al. 1985) ou d'électrophorèse capillaire.

Elle possède la caractéristique de produire des ions multichargés. En effet, les différents états de charge observés pour un peptide ou une protéine proviennent à la fois de la charge nette qu'ils possèdent lorsque le pH de la solution est différent de leur point isoélectrique, et à la fois de l'ionisation que provoque le processus d'électrospray. Des composés de plus en plus lourds ainsi que des complexes ont ainsi pu être analysés, du fait d'un rapport m/z analysé compatible avec les gammes de masse de la plupart des spectromètres de masse.

L'ionisation de type ESI peut être divisée en trois étapes (Kearle 2000): la formation de gouttelettes chargées à partir de la solution d'analytes, la production de gouttelettes filles et la production des ions en phase gazeuse (figure 8).

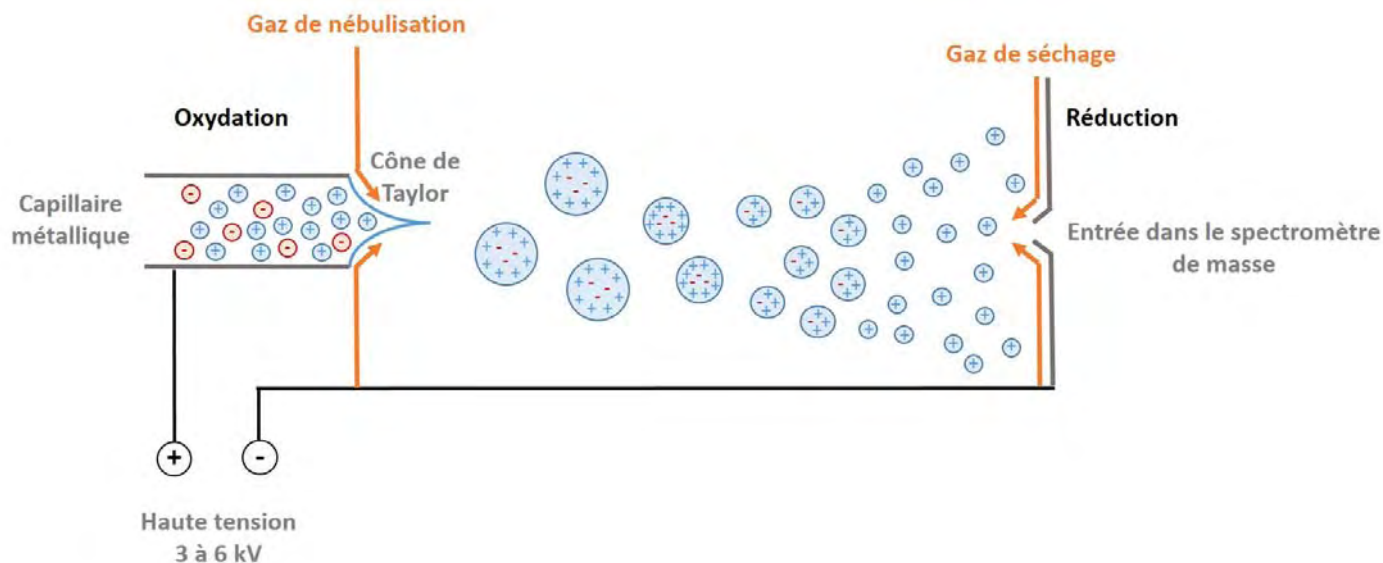


Figure 8: Principe de l'ionisation electrospray en mode positif

En mode d'ionisation positif, une différence de potentiel de 3 à 6 kV est appliquée entre un capillaire métallique contenant la solution d'analytes et une contre-électrode (entrée du spectromètre de masse). Cette différence de potentiel va entraîner une séparation électrophorétique des espèces de charges opposées au sein du liquide : les ions de charges positives sont repoussés et se concentrent en surface, tandis que les ions de charges négatives se trouvent au cœur du liquide. Le liquide est alors déstabilisé et un cône de Taylor se forme à l'extrémité du capillaire. Des gouttelettes multichargées sont émises et un flux d'azote coaxial, le gaz de nébulisation (N₂ le plus couramment), permet de les stabiliser. Le chauffage de l'enceinte aide à l'évaporation du solvant. Le volume des gouttelettes chargées va diminuer jusqu'à la limite de stabilité de Rayleigh. Les forces de répulsion des ions vont devenir supérieures à celles de la tension de surface, et vont entraîner une explosion coulombienne. Deux modèles décrivent la production des ions désolvatés et complètement libres. Le premier modèle, celui de Dole (Dole, Mack et al. 1968) propose une évaporation du solvant et des explosions coulombiennes qui se répètent jusqu'à ce qu'il n'y ait qu'une charge par gouttelette. Le Modèle d'Iribarne et Thomson, lui, propose le modèle selon lequel, lors du processus d'évaporation du solvant, la densité de charge est suffisante pour émettre directement les ions en phase gazeuse (Thomson and Iribarne 1979).

La source ESI conventionnelle nécessite des débits entre 1 et 100 $\mu\text{L}/\text{min}$. En 1996, des sources fonctionnant avec des débits beaucoup moins importants apparaissent (20 nL/min en infusion directe). C'est le début des sources nano-electrospray (Wilm and Mann 1996). Avec ces nouveaux débits, le diamètre de l'orifice de sortie de la source est diminué, l'utilisation du gaz de nébulisation n'est plus nécessaire et la taille des gouttelettes formées est plus petite, ce qui permet une meilleure efficacité d'ionisation et donc un gain de sensibilité.

2. L'analyse de type « bottom-up »

a) Le principe

Au cours de l'analyse d'un peptide, la simple mesure d'un m/z ne permet pas de déterminer l'enchaînement en acides aminés (les peptides AADD et ADAD, isobares, ont par exemple le même m/z). La fragmentation en tandem ou fragmentation MS/MS permet alors de rompre les liaisons peptidiques entre les acides aminés en fragments identifiables pour déterminer l'enchaînement des acides aminés.

Cette fragmentation MS/MS peut se faire sur les protéines, la méthode est alors nommée « top-down », ou sur les peptides avec une méthode « bottom-up » (figure 9). Les peptides sont alors issus d'une digestion enzymatique des protéines.

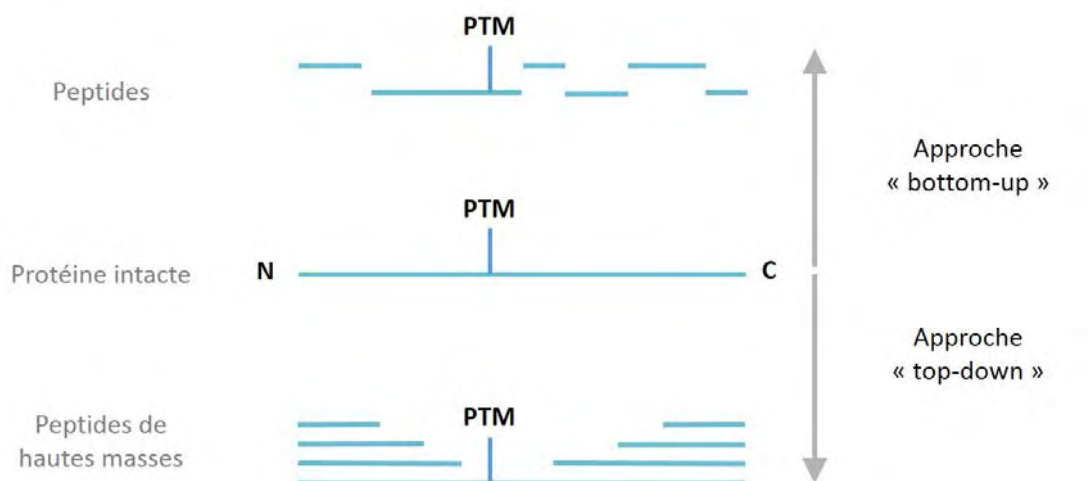


Figure 9: Les deux approches “bottom-up” et “top-down” pour l’analyse des protéines par spectrométrie de masse

La fragmentation des peptides et des protéines peut se faire par différents modes de fragmentation qui seront détaillés par la suite.

L'approche de type « bottom-up » est la stratégie de choix pour la définition d'un protéome, la recherche et l'identification d'un grand nombre de protéines au sein de mélanges complexes. Afin d'avoir une image exhaustive des protéines présentes dans un échantillon d'intérêt, la protéomique a bénéficié au cours des dernières années d'avancées à plusieurs niveaux, et notamment au niveau instrumental avec des spectromètres de masse qui possèdent de grandes vitesses de balayage et permettent donc l'acquisition d'un grand nombre de données, avec une sensibilité satisfaisante et de bonnes résolutions. Enfin, afin de retraiter toutes les données acquises au cours de telles analyses, des logiciels ont été

développés et sont sans cesse améliorés afin d'obtenir des identifications de protéines rapides et fiables.

Une analyse de type MS/MS peut être schématisée en trois étapes (figure 10).

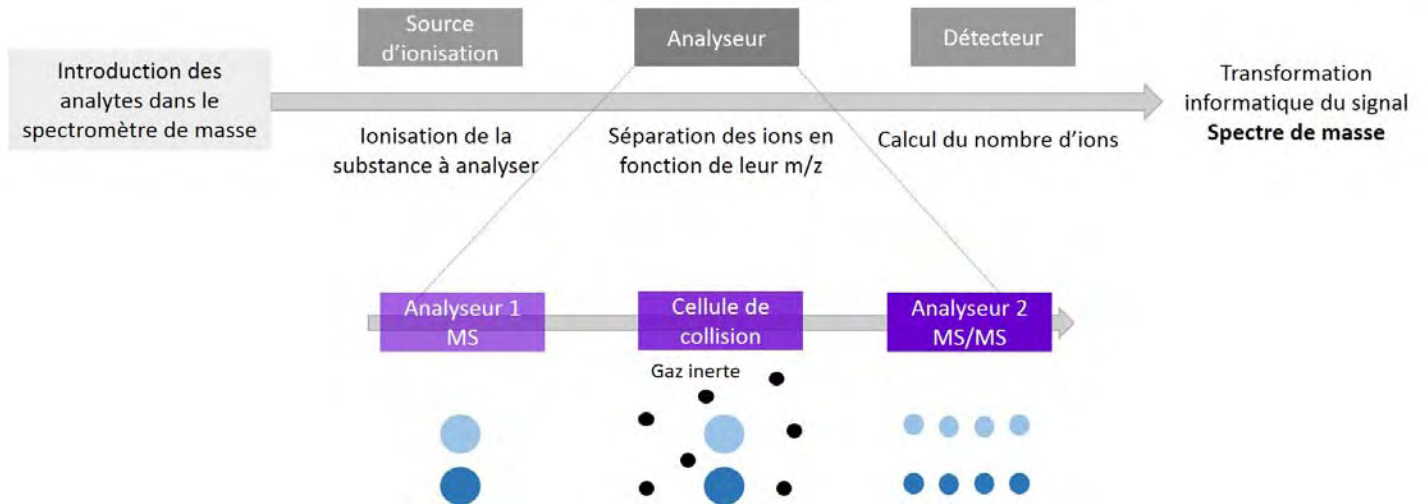


Figure 10: Les différentes étapes d'une analyse de type MS/MS

Au cours d'une première étape, un ion précurseur d'un m/z donné est sélectionné de façon spécifique dans un premier analyseur du spectromètre de masse.

L'ion est ensuite transféré dans une cellule de collision dans laquelle il va être fragmenté. La méthode la plus couramment utilisée pour apporter au peptide l'énergie nécessaire à sa fragmentation est la dissociation induite par collision (CID pour « collision-induced dissociation »).

b) Le mode de fragmentation CID

Dans la cellule de collision, l'ion percute un flux d'atomes de gaz inerte (azote, hélium ou argon), et l'énergie cinétique apportée au peptide est alors transformée en partie en énergie vibrationnelle. Cette énergie interne va se répartir de façon uniforme sur le squelette peptidique, ce qui conduit à la rupture des liaisons les plus labiles. Cette fragmentation intervient principalement au niveau des liaisons amides entre les acides aminés, ce qui génère des séries de fragments N-terminaux et C-terminaux appelés ions b et ions y dans la nomenclature conventionnelle (figure 11), d'après les travaux de Biemann (Biemann 1990).

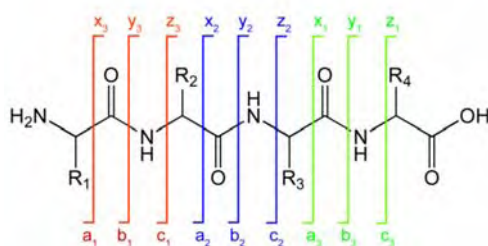


Figure 11: La nomenclature de Biemann des différents ions formés au cours d'une fragmentation CID (Biemann,1990)

Pour expliquer cette fragmentation, le modèle le plus admis est le modèle du proton mobile: les ions activés par une faible énergie de collision subissent une fragmentation dirigée par la charge (Cox, Gaskell et al. 1996). Certains sites du peptide sont favorables à la protonation (amine N-terminale, fonctions basiques des chaînes latérales), ou moyennement ou peu favorables (atomes d'oxygène et d'azotes des liaisons amides).

Si l'on considère un peptide dichargé, une des charges est « séquestrée » sur la lysine ou l'arginine C-terminale, et l'autre proton est appelé le « proton mobile ». Les collisions des peptides avec les molécules de gaz vont entraîner la migration de ce proton des sites les plus favorables vers les sites les moins favorables.

Les fragments générés sont ensuite analysés dans un second analyseur du spectromètre de masse, d'où le terme de spectrométrie de masse en tandem. Le détecteur mesure ensuite l'intensité de chacun des fragments produits et un spectre de masse appelé MS/MS ou MS² est généré. Une analyse bio-informatique de tous les spectres MS de premier niveau et de tous les spectres MS/MS de second niveau, avec des recherches en banques de données permet de remonter aux informations de séquence de peptides puis des protéines.

3. L'appareillage utilisé

Au cours de ma thèse, j'ai utilisé deux appareils de type LTQ Orbitrap Velos du constructeur Thermo dont l'un d'entre eux possède le module de fragmentation ETD.

Le LTQ Orbitrap Velos est un spectromètre de masse hybride (constitué de deux analyseurs) avec une source d'ionisation ESI, une trappe linéaire et un analyseur Orbitrap.

L'Orbitrap est un analyseur utilisant la transformée de Fourier, fonctionnant avec un champ électrostatique, présenté pour la première fois à la conférence annuelle de l'ASMS (American Society for Mass Spectrometry) en 1999.

Il est composé de deux électrodes, une électrode centrale qui possède une forme de fuseau et une électrode axiale qui possède une forme de tonneau coupé en deux parties égales (figure 10). Un vide poussé règne dans l'Orbitrap (10⁻¹⁰ torr). Les ions entrent dans l'Orbitrap par l'interstice de séparation entre les deux moitiés de l'électrode axiale. Un champ

électrique quadri-logarithmique est maintenu entre les électrodes. Les ions oscillent en orbite autour de l'électrode centrale dans les directions axiales et radiales sous l'action concomitante du champ électrique et de la force centrifuge qui s'exerce sur les ions. A chaque passage devant l'électrode axiale, chaque ion génère un courant induit par son mouvement sur lequel est appliquée une transformée de Fourier afin d'obtenir l'information du m/z (Scigelova and Makarov 2006).

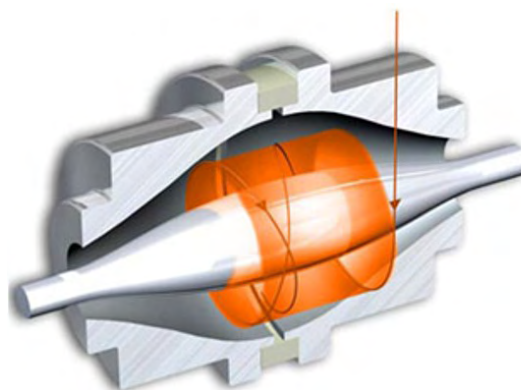


Figure 12: L'analyseur Orbitrap composé d'une électrode centrale et d'une électrode axiale

L'Orbitrap est un analyseur de masse d'une résolution de 100 000 à $m/z=400$, permettant une précision sur la mesure des m/z de l'ordre de 2 ppm au cours de l'analyse de mélanges peptidiques complexes (Yates, Cociorva et al. 2005). Cette performance peut être améliorée avec l'utilisation d'une calibration interne (Olsen, Schwartz et al. 2009).

- **Le mode d'acquisition MS**

Les ions provenant de la source ESI sont transmis dans la trappe linéaire (partie 5, figure 13) dans laquelle ils sont piégés. Les ions sont ensuite envoyés vers la C-trap (« curved trap » partie 7, figure 13), dans laquelle ils perdent leur énergie cinétique en entrant en collision douce avec des molécules de gaz, sans subir de fragmentation.

Les ions sont ensuite envoyés par petits paquets de la C-trap vers l'Orbitrap (partie 11, figure 13), dans lequel ils sont détectés à très haute résolution.

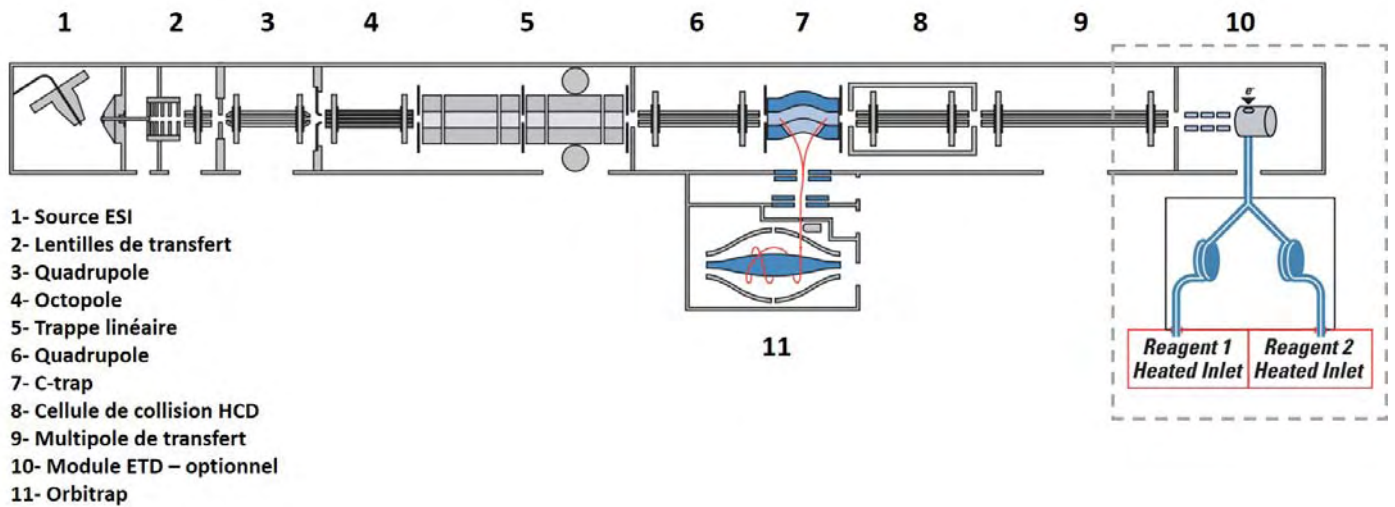


Figure 13: Les différentes parties d'un appareil LTQ-Orbitrap de Thermo

- **Le mode d'acquisition MS/MS**

Au cours d'une analyse classique de type « bottom-up » au laboratoire, le mode d'acquisition DDA (« data dependent acquisition ») est utilisé. Avec ce mode d'acquisition, un premier spectre MS basse résolution est acquis dans l'Orbitrap, le « pré-scan », afin de visualiser quels sont les ions présents au sein du spectromètre de masse à un instant t (Olsen, Schwartz et al. 2009). Parmi ces ions, les N ions les plus intenses (avec N souvent égal à 20), dont le signal dépasse un certain seuil d'intensité fixé par l'utilisateur, et dont l'état de charge est conforme aux critères fixés, sont sélectionnés pour être fragmentés dans la trappe linéaire. Des cycles de MS et MS/MS sont alors réalisés en parallèle : pendant l'analyse MS haute résolution dans l'Orbitrap des ions précurseurs, la trappe linéaire (LTQ, partie 5, figure 11), plus rapide (20 spectres en moins de 3 secondes), acquiert les spectres MS/MS des peptides les plus intenses préalablement définis par le pré-scan (figure 14).

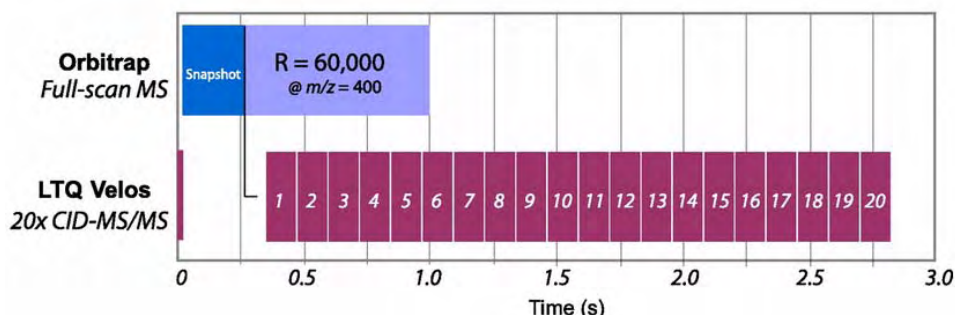


Figure 14: Les cycles de MS et MS/MS au cours d'une analyse de type DDA avec un appareil LTQ-Orbitrap Velos (fragmentation CID). En moins de 3s, 20 MS/MS sont enregistrées.

Figure de Olsen et al. 2009

Afin que ça ne soit pas toujours les mêmes ions intenses qui soient sélectionnés pour la fragmentation, une exclusion dynamique a été mise en place : lorsque un ion a été fragmenté, sa masse est entrée dans une liste d'exclusion et il ne peut plus être soumis à la fragmentation sur une durée définie par l'utilisateur.

La résolution maximale de l'Orbitrap est de 100 000 mais le temps requis pour l'acquisition d'un spectre MS est proportionnel à cette résolution. De façon pratique, une résolution de 60 000 dans l'Orbitrap est souvent choisie, ce qui permet l'acquisition d'un spectre MS en moins d'une seconde. Une acquisition haute résolution de la masse des ions précurseurs permet par la suite une interrogation plus spécifique des bases de données.

4. Cas particulier des peptides glycosylés

D'autres modes de fragmentation existent en spectrométrie de masse, qui peuvent être intéressants dans le cas de l'étude de peptides portant certaines modifications post-traductionnelles.

En effet, au cours de la fragmentation CID d'un peptide portant une PTM labile (glycosylation par exemple), l'énergie vibrationnelle apportée au peptide peut, dans un premier temps, entraîner la rupture de la liaison entre cette modification et le squelette peptidique avant même de permettre la formation des ions y et b (figure 15, Scott, Parker et al. 2011).

L'information sur la localisation de la modification est perdue et d'autres modes de fragmentation permettent alors d'obtenir cette information et peuvent être utiles pour l'étude de toutes les PTMs labiles, comme la glycosylation.

Pour une recherche de type « bottom-up » classique dédiée à l'identification globale de protéines, la fragmentation CID est la plus souvent utilisée mais d'autres types de fragmentation peuvent être utiles pour la recherche et la caractérisation de glycopeptides.

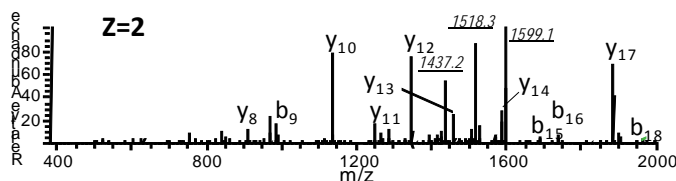


Figure 15: Spectre de masse d'un peptide tri-glycosylé **dichargé**. Au cours de la fragmentation du peptide avec un mode de fragmentation CID, il y a formation au sein du spectre de 3 ions provenant des 3 pertes d'hexoses successives à partir de l'ion parent (ions soulignés). Ce spectre MS/MS possède alors une signature caractéristique de peptide glycosylé.

a) La fragmentation HCD

Parmi les autres types de fragmentation, la fragmentation HCD (Higher-energy Collisional) (Olsen, Macek et al. 2007) est une variante de la fragmentation CID mais avec une énergie apportée aux peptides qui est plus importante, et conduit à la formation des mêmes ions y et b, avec une prépondérance d'ions y.

Sur l'Orbitrap Velos, la fragmentation des peptides a lieu dans un octopole placé en aval de la C-trap (partir 8, figure 13). Les fragments sont ensuite envoyés vers la C-trap puis enfin vers l'Orbitrap dans lequel ils sont détectés.

Les spectres MS/MS issus d'une fragmentation HCD contiennent des ions fragments de bas poids moléculaires, et peuvent donc être informatifs au cours d'une recherche de modifications post-traductionnelles. De plus, du fait de la détection des fragments dans l'Orbitrap, une précision de masse satisfaisante sur chacun des fragments ainsi qu'une meilleure résolution sont obtenues, ce qui permet d'augmenter la qualité de l'interrogation des bases de données.

En revanche, la détection de l'ion précurseur mais aussi des ions fragments se faisant dans l'Orbitrap, une détection en parallèle du précurseur et des fragments comme il est possible avec la fragmentation CID n'est pas possible dans le cas de la fragmentation HCD (figure 16). Les temps de cycle avec fragmentation HCD sont donc plus longs. Selon le type d'analyse que l'utilisateur doit mener, il doit alors faire un choix entre les avantages qu'apporte la fragmentation HCD avec détection des fragments dans la trappe et le nombre moins importants de spectres MS/MS enregistrés au cours d'une analyse à haut-débit.

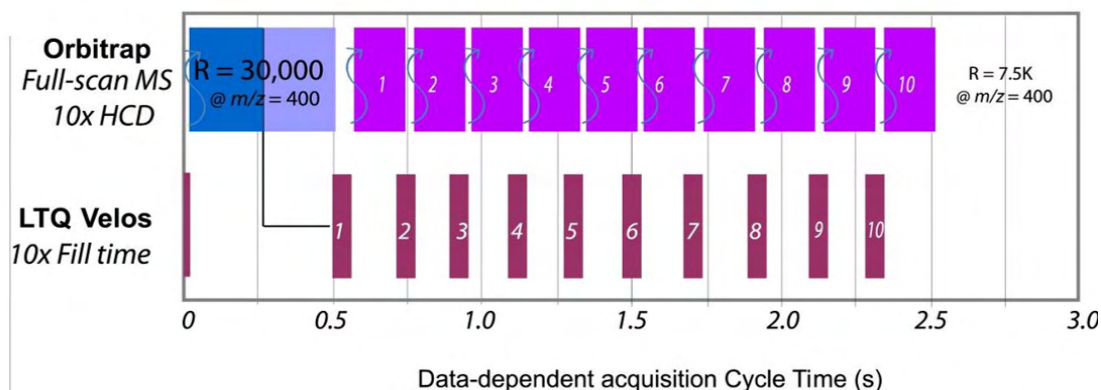


Figure 16 : Au cours de la fragmentation HCD, les cycles MS et MS/MS se sont de façon alternative. En moins de 3s, 10 MS/MS sont enregistrées.

Figure de Olsen et al. 2009

b) La fragmentation ETD

La fragmentation ETD (Electron Transfer Dissociation) est décrite par Syka *et al.* en 2004 (Syka, Coon et al. 2004). Au cours de cette fragmentation, il y a transfert d'un électron à partir d'un radical anion (provenant d'une molécule présente dans la source ETD, le fluoranthène par exemple) vers le peptide (figure 17). De la même façon que pour la fragmentation ECD, ceci induit la fragmentation du peptide et la formation d'ions fragments de type c et z. La fragmentation ETD nécessite une trappe ionique et non un FT-ICR pour la sélection des ions et la détection.

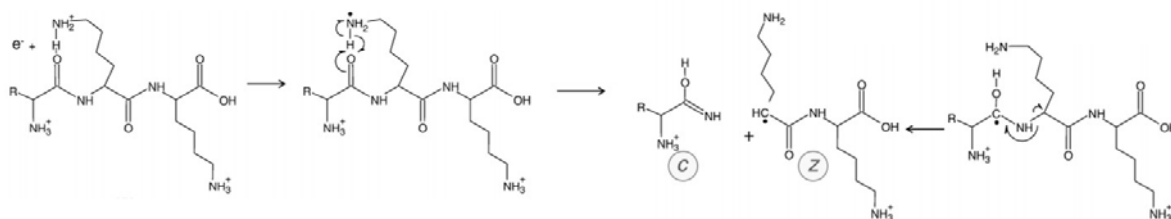


Figure 17: Le mécanisme de fragmentation ETD selon Syka *et al.*

Avec ce mode de fragmentation, les modifications post-traductionnelles labiles ne sont plus éliminées du peptide. En effet, contrairement à la fragmentation CID pour laquelle c'est une énergie apportée au peptide qui entraîne sa fragmentation, le mode de fragmentation ETD qui fonctionne par un transfert d'un électron sur la structure du peptide permet de conserver de telles modifications. La fragmentation ETD est particulièrement intéressante

pour la détermination des glycosites (Steentoft, Vakhrushev et al. 2011), ou des phosphosites par exemple (Frese, Zhou et al. 2013).

De plus, en 2008, Swaney et al. mènent une étude afin de déterminer quelle est la fragmentation la plus adaptée selon la masse et la charge des ions précurseurs (figure 18) (Swaney, McAlister et al. 2008). La fragmentation ETD semble alors apporter une bonne fragmentation des ions possédant de faibles m/z , alors que la fragmentation CID apporte une meilleure fragmentation des ions possédant de moyens à hauts m/z . Pour les peptides di-chargés, la fragmentation CID est toujours plus favorable que la fragmentation ETD. Il est alors possible d'utiliser un algorithme qui permet une sélection automatique et en temps réel du mode de fragmentation le plus adapté selon la masse et la charge de l'ion sélectionné pour la fragmentation, c'est le mode d'acquisition « data-dependent decision tree » (figure 18).

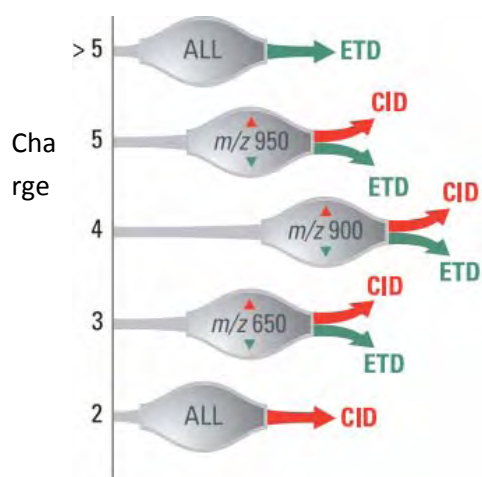


Figure 18: Le type de fragmentation préférentiel selon la charge et la masse de l'ion précurseur, d'après la méthode « data dependent decision tree »

Enfin, la fragmentation ETD semblerait produire des spectres plus informatifs que le mode de fragmentation CID (un nombre plus importants d'ions correspondant aux fragments), permettant une meilleure attribution de séquence peptidique (Mikesh, 2006). En effet, avec une fragmentation CID, lorsque le peptide contient par exemple de nombreux résidus arginines, ceux-ci vont gêner la migration du proton mobile le long de la séquence peptidique et donc gêner la fragmentation. La fragmentation ETD permet de s'affranchir de ce problème et les spectres MS/MS obtenus sont alors plus complets avec des ions fragments couvrant l'ensemble de la séquence peptidique.

L'analyse nanoLC-MS/MS des peptides génère une grande quantité de données MS et de données MS/MS. Au regard du volume considérable de données à traiter, de nombreux

moteurs de recherche ont été développés permettant d'automatiser l'identification de peptides et de protéines à partir des spectres MS et MS/MS.

C. Le retraitement bio-informatique des données

I. Cas général

Plusieurs méthodes ont été développées afin de réaliser l'identification de protéines à partir de spectres MS/MS (Nesvizhskii 2010).

L'approche par empreinte de fragmentation peptidique (PFF pour « Peptide Fragment Fingerprinting ») est basée sur la comparaison des listes de masses des ions parents et de leurs fragments associés (extraites des spectres MS et des spectres MS/MS respectivement) avec des listes de masses issues de la digestion théorique des protéines *in silico* présentes dans des banques de séquences protéiques, et de la fragmentation théorique des peptides associés (figure 19). Au laboratoire, cette comparaison est principalement réalisée grâce au moteur de recherche Mascot (Perkins, Pappin et al. 1999). D'autres logiciels utilisent également cette approche, comme OMSSA (Geer, Markey et al. 2004).

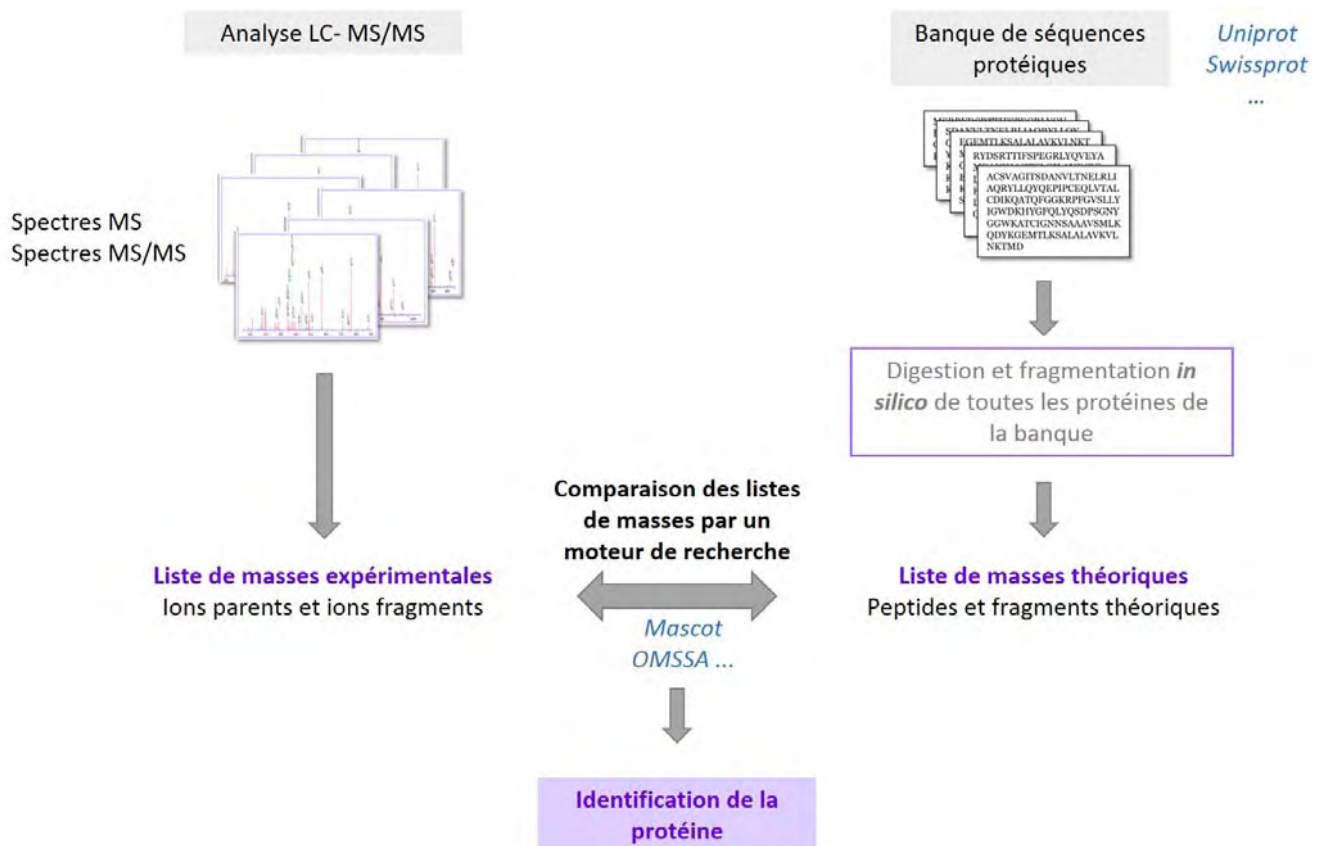


Figure 19: Les différentes étapes permettant l'identification d'une protéine à partir des données LC-MS/MS brutes

Tous ces logiciels fonctionnent à partir de paramètres que définit l'utilisateur : l'erreur de masse tolérée sur l'ion parent et sur les ions fils, l'enzyme de digestion utilisée (coupures spécifiques ou non spécifiques), les modifications des acides aminés, le type de série de fragmentation recherchée et le format de la liste de masses.

En revanche, la différence entre ces différents logiciels réside dans la fonction de corrélation utilisée pour déterminer le degré de similarité entre un spectre MS/MS théorique et un spectre MS/MS expérimental, et donc dans la vraisemblance de l'identification. Le logiciel Mascot associe une valeur de probabilité à l'attribution d'une séquence peptidique à un spectre MS/MS donné, elle est basée sur une implémentation probabilistique de l'algorithme MOWSE (Perkins, Pappin et al. 1999). Ce score de corrélation est défini par la formule :

$$\text{Ion score} = -10 \log_{10}(P)$$

où P représente la probabilité qu'une attribution peptidique donnée soit due au hasard. Ce score est donc représentatif de la fiabilité de l'attribution.

L'utilisation de cette approche pour déterminer la séquence des peptides peut entraîner de fausses attributions.

En effet, cette stratégie n'est adaptée que pour les protéines présentes dans les banques de séquences protéiques. Considérons un spectre MS/MS correspondant réellement à un peptide n°1 d'une protéine qui n'est pas présente dans la banque de données que l'on interroge. Si certains des fragments de ce peptide sont communs à ceux d'un autre peptide n°2 contenu dans la base de données, mais qui ne correspond pas réellement au spectre MS/MS étudié, alors la séquence du peptide n°2 sera attribuée au spectre MS/MS mais cette interprétation sera fausse. Il faut être alors vigilant à la base de données que l'on utilise.

De plus, lorsque le spectre est de mauvaise qualité et peu informatif, l'attribution peptidique peut également être fausse. Enfin, de fausses interprétations peuvent également être obtenues lorsque les paramètres de recherche choisis par l'utilisateur ne sont pas assez larges.

Avec le logiciel Mascot, une valeur de seuil d'identité est calculée, pour chaque interprétation peptidique et il est ainsi possible d'évaluer la fiabilité de chacune d'entre elles. Elle est définie par la formule :

$$\text{Identity threshold} = -10 \log_{10} \left(\frac{p}{N} \right)$$

où p est le seuil de significativité (p-value) et N correspond au nombre de peptides candidats.

Le seuil d'identité reflète ainsi le niveau de significativité que l'on veut se fixer (p-value) et également la taille de l'espace de recherche. Plus l'espace de recherche est important, plus le risque de faux-positifs est élevé. La taille de cet espace est défini par la taille de la banque

de données, à la spécificité de l'enzyme, aux modifications post-traductionnelles, et enfin à la valeur de la tolérance de masse choisie pour l'ion précurseur.

D'autre part, des méthodes ont été développées afin de filtrer les interprétations peptidiques obtenues et d'évaluer le taux « faux-positifs » ou FDR. La stratégie la plus connue pour cela est la stratégie de « target decoy » (Elias and Gygi 2010). Dans cette stratégie, la recherche d'identification est réalisée dans la banque protéique d'intérêt, la banque « target », et également dans une banque « decoy » qui contient toutes les séquences inversées. Les résultats d'identification contiennent alors à la fois des vrais et des faux positifs. Les identifications des peptides sont alors filtrées afin d'obtenir un taux de faux positifs le plus bas possible.

Une fois l'identification des peptides réalisée, l'étape suivante au cours d'une analyse « bottom-up » pour l'identification des protéines au sein d'un échantillon est l'étape d'assemblage des séquences peptidiques aux protéines contenues dans la base de données. Cette tâche peut être complexe puisque un ensemble de séquences peptidiques peut correspondre à plusieurs protéines alors homologues ou isoformes, ou encore à plusieurs protéines possédant des portions de séquences identiques dans la base de données. De plus, un peptide peut également correspondre à plusieurs séquences protéiques.

L'approche utilisée par le moteur de recherche Mascot est une méthode dite « parcimonieuse » qui consiste à trouver la liste de protéines la plus petite capable d'expliquer l'ensemble des séquences peptidiques identifiées. Ainsi, l'ensemble des numéros d'accèsion protéiques associés exactement au même ensemble de peptides (peptides « same-set ») sont regroupés et considérés comme une identification. De plus, des numéros d'accèsion qui aurait pu être associés uniquement à un sous-ensemble des mêmes peptides sont éliminés (peptides « sub-set »).

II. Cas des protéines glycosylées

Selon le type de glycosylation, différentes stratégies de glycoprotéomique peuvent être choisies. Dans le cas d'une N-glycosylation, il peut être choisi soit d'analyser les glycopeptides intacts par spectrométrie de masse (stratégie « one step »), soit de procéder à une étape de déglycosylation des glycopeptides avant l'analyse (stratégie « two-steps ») (Li, Glinskii et al. 2013). Dans ce dernier cas, les peptides glycosylés sont analysés d'une part, et les glycannes libérés le sont d'autre part. Dans le cas d'une déglycosylation d'un glycopeptide avec la PNGase par exemple, l'asparagine porteuse du glycane est transformée en acide aspartique, ce qui entraîne un gain de masse de 1 Da sur le peptide, correspondant à une déamidation. Il est alors possible d'identifier les peptides déglycosylés par une recherche de cette déamidation.

Dans le cas d'une *O*-glycosylation, pour laquelle il n'existe pas de glycosidase spécifique, des méthodes chimiques permettent la déglycosylation d'un *O*-glycopeptide mais j'ai choisi au cours de ma thèse de ne pas utiliser ces méthodes et d'analyser des glycopeptides intacts. En effet, j'ai choisi de mettre en place une stratégie de recherche des glycopeptides/glycoprotéines qui ne fasse intervenir aucune étape de préparation d'échantillon supplémentaire par rapport à une stratégie protéomique de type « bottom-up » classique.

Afin d'identifier les *O*-glycopeptides séquencés au cours d'une analyse de glycoprotéomique, il est alors nécessaire d'utiliser des programmes bioinformatiques adaptés pour une stratégie « one-step ».

1. Méthodes de glycoprotéomiques nécessitant une analyse préliminaires des protéines de l'échantillon

A partir des données d'acquisition brutes acquises au cours de l'analyse des échantillons d'intérêt, renfermant des peptides non-glycosylés ainsi que des peptides glycosylés intacts, certaines des méthodes de recherche des glycopeptides comportent une première étape de recherche et de validation globale des protéines contenues dans l'échantillon. La recherche des attributions peptidiques des spectres MS/MS dans les banques de données est réalisée sans aucune PTM, seuls des peptides non modifiés sont recherchés. Une banque de données réduite est ensuite créée, renfermant la liste des protéines identifiées et validées.

- **Recherche de PTMs « à l'aveugle » par une méthode d'alignement de spectres (Tsur, Tanner et al. 2005, Chalkley, Baker et al. 2008).**

Sur un ensemble de spectres MS/MS, une recherche de PTMs « à l'aveugle » peut être effectuée sans à priori sur les PTMs portées par les peptides. Une recherche des protéines contenues dans l'échantillon est effectuée au préalable (cf. paragraphe précédent). Une corrélation est ensuite recherchée entre les pics présents au sein des spectres MS/MS des peptides de ces protéines validées et les spectres MS/MS de spectres encore non interprétés. Dans le cas de deux spectres MS/MS partageant une même signature de spectre (des ions y et b en commun), et dont les ions parents diffèrent d'une masse m , une PTM (ou une résultante de PTM) de masse m peut alors être envisagée.

Grâce à cette méthode, une recherche exhaustive de PTMs au sein d'un jeu de données peut être effectuée. Cependant, une telle recherche n'est possible que sur un jeu de données réduit. En effet, l'espace de recherche contenant les spectres MS/MS de référence, pour lesquels des spectres MS/MS possédant des ions fragments en commun vont être recherchés, ne doit pas être d'une taille trop importante, afin d'éviter des temps de retraitement des données trop importants ou éviter que les ressources informatiques

utilisées ne soient pas assez puissantes. En travaillant avec une liste de protéines validées au cours d'une étape préliminaire, l'utilisateur accepte que seuls des peptides modifiés associés à des peptides de cette liste de protéines réduite soient identifiés.

- **Digestion *in silico*** d'une banque théorique de glycopeptides (Wang, Wong et al. 2011)

En 2011, Wang *et al.*, dans Nature Protocols, ont décrit la méthode IGAP (Intact Glycopeptide Analysis Pipeline) pour une recherche automatique de *N*-glycopeptides.

Cette méthode implique l'utilisation de deux bases de données, l'une renfermant les protéines que l'on souhaite étudier, l'autre contenant les *N*-glycannes d'intérêt, susceptibles d'être portés par les protéines contenues dans la première base de données.

Une digestion *in silico* des peptides tryptiques de la banque de données protéique, d'une part, est réalisée (avec un nombre de clivages manqués maximum de 2). Seuls les peptides possédant la séquence consensus N-X-S-T (X différent d'une proline) connue pour la *N*-glycosylation sont considérés comme *N*-glycopeptides potentiels. D'autre part, une digestion théorique des glycannes est effectuée (ions Y et B selon la nomenclature de Domon et Costello (Domon and Costello 1988)). Les données issues des deux digestions théoriques sont réunies afin de générer des spectres MS/MS théoriques. Une comparaison des spectres MS/MS expérimentaux et des spectres MS/MS théoriques est alors effectuée. Des valeurs de scores sont attribuées aux interprétations des spectres MS/MS.

L'existence d'une séquence consensus pour la *N*-glycosylation permet de filtrer les spectres MS/MS théoriques de glycopeptides potentiels et de réduire le nombre de ces spectres à comparer aux spectres expérimentaux.

En revanche, dans le cas d'une *O*-glycosylation, aucune séquence consensus n'est connue afin de prédire les sites de glycosylation. Pour une recherche de glycopeptides à grande échelle, la liste des spectres MS/MS théoriques de peptides glycosylés à comparer à la liste de spectres expérimentaux peut alors être relativement importante et limitante.

2. Méthodes de glycoprotéomiques ne nécessitant pas d'analyse préliminaires des protéines de l'échantillon

- **Recherche d'un ion oxonium diagnostique** (Carapito, Klemm et al. 2009)

Au cours d'une fragmentation de type CID ou HCD d'un peptide glycosylé (ou phosphorylé également), un ion oxonium diagnostique étant formé (provenant de la rupture de la liaison glycosidique), la recherche de cet ion diagnostique au sein de tous les spectres MS/MS acquis peut être une méthode permettant de rechercher les glycopeptides. En revanche, afin de pouvoir visualiser ces ions au sein des spectres de masse, il est nécessaire que la

plage de masse d'acquisition soit suffisamment large, notamment dans les basses masses. Sur l'Orbitrap velox, ceci n'est possible que dans le cas de la fragmentation HCD pour laquelle les ions fragments sont détectés dans l'Orbitrap. Or, comme cela a déjà été dit, les temps de cycle sont alors plus longs et le nombre de spectres MS/MS enregistré est alors plus faible.

- **Recherche de perte(s) de neutre(s)**

Egalement au cours d'une fragmentation CID ou HCD d'un peptide glycosylé, un ion intense provenant de la perte de la modification à partir de l'ion moléculaire étant formé, il est également possible de rechercher cet ion caractéristique au sein des spectres MS/MS. Smith *et al* en 2013, en recherchant ce type de spectres au sein de données issues de l'analyse des protéines du surnageant de culture de *Mycobacterium tuberculosis* et en utilisant d'autres méthodes analytiques, ont pu identifier 13 nouvelles glycoprotéines de *Mycobacterium tuberculosis* (Smith, Sweredoski et al. 2014).

Au début de ma thèse, aucun programme permettant la recherche à grande échelle de spectres MS/MS possédant cet ion caractéristique n'avait encore été développé. Au sein du laboratoire, nous avons alors développé un tel programme dédié à la recherche des glycopeptides.

D. Les méthodes d'analyse de protéines entières pour la caractérisation des modifications post-traductionnelles

La stratégie d'identification des protéines de type « bottom-up » permet une identification des peptides modifiés, ainsi que la localisation des sites de modification grâce à l'analyse MS/MS. Or, au cours d'une analyse de ce type, il est bien admis qu'une proportion non négligeable de peptides n'est pas identifiable (peptides de trop faible masse, peptides très hydrophobes etc (Eyers, Lawless et al. 2011)), et les informations concernant d'éventuelles PTMs sur ces peptides ne peuvent être atteintes.

Par une analyse de protéines entières par spectrométrie de masse, il peut alors être possible d'obtenir des informations complémentaires à celles obtenues via une analyse de type « bottom-up », notamment la nature de l'ensemble des PTMs de chaque protéine. De plus, la fragmentation d'une protéine entière selon une stratégie « top-down » et avec le mode de fragmentation ETD peut permettre de localiser les sites de modification et de définir la connectivité entre celles-ci.

I. L'analyse de protéines entières par spectrométrie de masse pour l'identification des modifications

L'analyse de protéines entière par spectrométrie de masse a connu un essor depuis le développement de méthodes de séparation de protéines et l'apparition d'analyseurs possédant de très hautes résolutions, qui permettent alors une détermination précise des hautes masses. La protéine groEL de *E. coli*, d'une masse de plus de 800 KDa a ainsi pu être analysée et détectée dans un Orbitrap avec une bonne précision de masse (Rose, Damoc et al. 2012).

De plus, l'analyse de protéines entières peut permettre une détermination rapide de la nature des PTMs. En effet, contrairement à l'approche de type « bottom-up », aucune étape de digestion enzymatique n'est nécessaire. Il est par exemple possible, en une seule analyse, de déterminer la nature des N-glycannes portées par un anticorps monoclonal de 150 KDa (Fornelli, Damoc et al. 2012), ou encore de détecter des isoformes dont seuls les ponts disulfures diffèrent (Bondarenko, Second et al. 2009)

1. Fractionnement des protéines

Dans le cas de l'étude d'un mélange de protéines modifiées ou de plusieurs isoformes d'une protéine d'une séquence donnée, il est indispensable d'utiliser une méthode de fractionnement en amont de l'analyse. En effet, il est nécessaire de réduire la complexité de l'échantillon à analyser ou de séparer les différentes formes d'une protéine pouvant porter différentes combinaisons des mêmes PTMs.

Pour ce dernier point, la mobilité ionique apparait comme une technique particulièrement intéressante. En effet, elle permet la séparation de deux protéines ou peptides isomères possédant les mêmes PTMs mais à des positions différents, avec des conformations légèrement différentes, alors que le pouvoir résolutif de la chromatographie liquide dans ce cas est souvent insuffisant (Gilar, Jaworski et al. 2007, Shvartsburg, Creese et al. 2010).

Les PTMs portées par les protéines peuvent également modifier leur caractère hydrophile ou hydrophobe. La chromatographie HILIC par exemple peut alors également permettre la séparation de protéines (Garcia, Pesavento et al. 2007).

Enfin, en 2008, une nouvelle technique de séparation « off-gel » des protéines a été introduite (Tran and Doucette 2008). Dans cette technique, les protéines traversent une colonne de gel de polyacrylamide parcouru par un champ électrique dans laquelle elles vont être séparées selon leurs poids moléculaires. Elles sont ensuite éluées et collectées en fractions liquides, d'où l'appellation de technique « off-gel » (Ahlf, Thomas et al. 2013). En utilisant une colonne de gel d'un pourcentage d'acrylamide de 12%, la plage de séparation se situe entre 3.5 et 50 kDa, avec une résolution entre 10 et 50 kDa. Utile pour séparer un mélange de protéines, cette technique ne semble en revanche pas permettre la séparation des différentes formes des protéines modifiées, en raison de la faible résolution obtenue.

2. L'introduction des protéines dans le spectromètre de masse

En amont de l'analyse par spectrométrie de masse, il est possible d'utiliser un couplage avec la chromatographie liquide ou d'injecter la protéine ou le mélange de protéines en infusion directe dans le spectromètre de masse.

Dans le cas d'un couplage avec la chromatographie liquide, de la même façon que pour l'analyse des peptides, la chromatographie en phase inverse est la technique la plus couramment utilisée. Les phases les plus utilisées sont des phases de type C4, C5, mais aussi C8 pour les protéines de faibles masses. Des phases polymériques peuvent également être utilisées pour l'analyse de protéines sur une plus vaste plage de masse (Wynne, Edwards et al. 2010). En revanche, le temps d'élution de la protéine peut être relativement court, entraînant l'acquisition au sein du spectromètre de masse d'un faible nombre de spectres. Il est alors possible que trop peu de données soient acquises pour obtenir une couverture de séquence de la protéine qui soit correcte (Lanucara and Eyers 2013).

II. La stratégie d'analyse de type « top-down » pour la localisation des modifications post-traductionnelles

Récemment, le terme « proteoform » a été introduit pour désigner les différentes formes (avec les différentes PTMs et leur différentes positions) d'une protéine donnée (Smith and Kelleher 2013). Caractériser les protéoformes d'une protéine consiste alors à rechercher les différentes formes d'une protéine dont les PTMs tout en gardant la connectivité entre celles-ci (figure 20). Considérons par exemple une protéine pour laquelle 3 PTMs ont été identifiées. Si la PTM 1 et la PTM 2 coexistent toujours sur la protéine et alors que la PTM 3 est toujours présente de façon isolée sur la protéine, il existe alors bien une connectivité particulière des PTMs de la protéine. Au cours d'une analyse « bottom-up », il est impossible de définir la connectivité des PTMs sur une protéine, seule la stratégie « top-down » permet alors d'obtenir cette information (Moradian, Kalli et al. 2014).

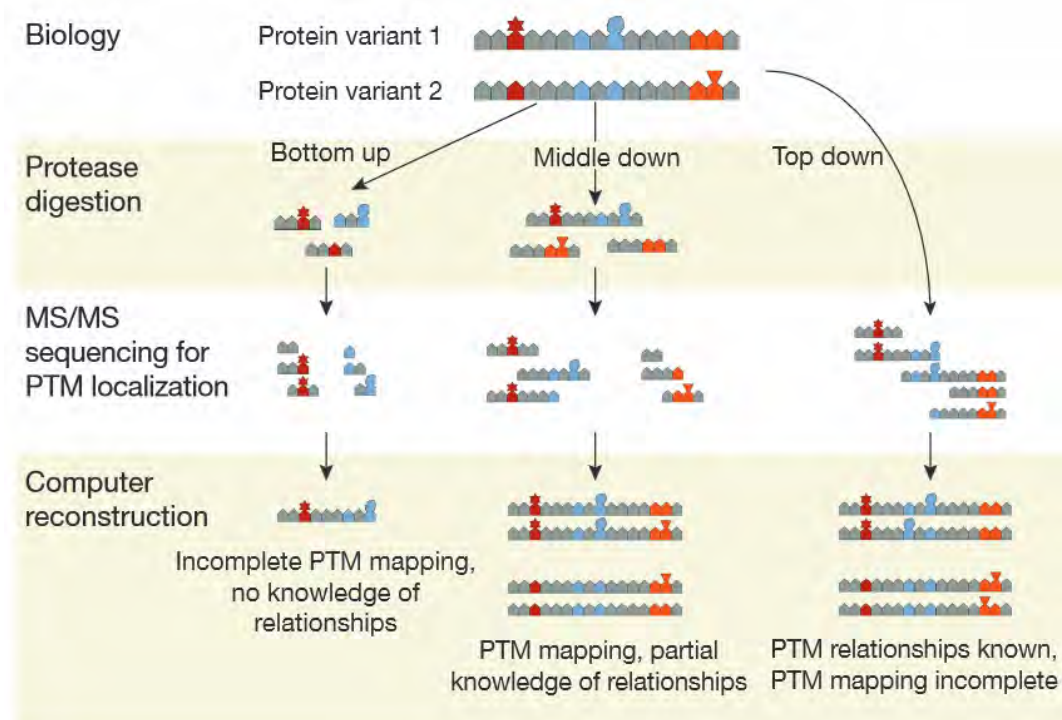


Figure 20 : Les informations que l'on peut atteindre sur les PTMs portées par deux protéines variantes (même séquence en acide aminés mais PTMs différentes), par trois approches : les approches « bottom-up », « top-down » et « middle-down ». L'approche « middle – down » consiste en la fragmentation et l'analyse de peptides de hautes masses (masses supérieures à 3 kDa) ((Moradian, Kalli et al. 2014)

1. La fragmentation des protéines

Un des principaux défis de l'analyse de protéines de type « top-down » est d'obtenir une fragmentation des protéines correcte afin d'obtenir une bonne couverture de séquence. Les modes de fragmentation qui sont bien connus sur les peptides peuvent conduire à des résultats très différents sur les protéines. En effet, leur efficacité dépend des propriétés intrinsèques des protéines, leur taille, leur charge, la densité de charge, et leur conformation en phase gazeuse.

Au cours de la fragmentation ETD, l'énergie apportée au peptide ou à la protéine permet le clivage des liaisons peptidiques avant toute redistribution de l'énergie sur toute la protéine via des énergies vibrationnelles. L'énergie apportée à la protéine est de l'ordre de quelques eV, alors que des différences d'énergie de dissociation selon les liaisons peptidiques de la protéine sont inférieures à 1 eV. Des ruptures préférentielles de certaines liaisons peptidiques aux dépiments d'autres sont alors minimisées et les couvertures de séquences obtenues grâce à la fragmentation ETD sont plus élevées que celles atteintes avec les fragmentations CID ou HCD (Coon, Ueberheide et al. 2005). De plus, il est possible d'utiliser différents temps d'activation entre la protéine et la molécule de fluoranthène. En effet, de

courts temps d'activation conduisent à l'obtention de nombreux fragments alors que des temps d'activation plus longs conduisent à des fragments aux extrémités de la protéine (Hao, Jiang et al. 2010).

Le mode de fragmentation en source SID (In source dissociation) peut également être utilisé sur les protéines, en augmentant par exemple la valeur de la tension en source. Ce mode de fragmentation conduit principalement à l'obtention de fragments sur les extrémités N- et C-terminales (Second, Zabrouskov et al. 2010) (figure 21). Peu d'indications sont données dans la littérature concernant le comportement des motifs hexoses portés par les protéines avec ce mode de fragmentation.

Enfin, le mode de fragmentation HCD utilisé sur les protéines permet l'obtention d'un nombre relativement faible de fragments, principalement au niveau des extrémités N- et C-terminales (Second, Zabrouskov et al. 2010) (figure 21). De la même façon que pour la fragmentation CID, peu d'indications sont données concernant l'étude des PTMs labiles avec ce mode de fragmentation, la fragmentation ETD étant la méthode de choix pour l'analyse des PTMs.

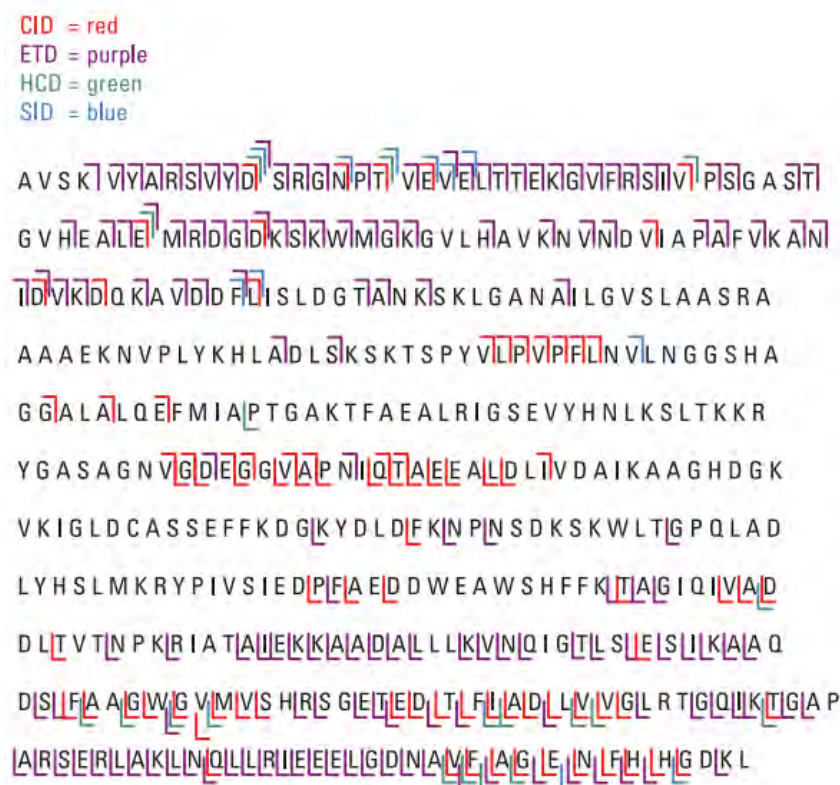


Figure 21: Les différents fragments de l'énolase obtenus avec différents modes de fragmentation. Note d'application 498 de Thermo Scientific.

2. Le retraitement bio-informatique des données

Les spectres acquis au cours d'une analyse « top-down » sont relativement complexes, avec des ions aux états de charge relativement différents. La première étape du retraitement bio-informatique est une étape de déconvolution de tels spectres pour obtenir une liste de masses monoisotopiques utilisée ensuite pour l'interprétation (Liu, Sirotkin et al. 2012).

Le logiciel ProSight PTM a été le premier outil bio-informatique disponible pour l'identification de protéines et la recherche de PTM (LeDuc, Taylor et al. 2004). Ce logiciel fonctionne avec une base de données dans laquelle toutes les PTMs sont répertoriées, ainsi que toutes les formes possibles d'une protéine d'une séquence donnée avec les combinaisons possibles de ces PTMs. Les données expérimentales sont ensuite confrontées à cette base de données théorique. La version commerciale ProSight PC a ensuite été développée par la société Thermo.

De plus, une version de Mascot dédiée à l'analyse de protéines a également été proposée, et permet de s'affranchir de la limite de masse sur le précurseur de 16 kDa sur la version traditionnelle.

Enfin, de la même façon que pour les analyses de type « bottom-up », des logiciels d'alignement de spectres permettent également le retraitement des données de type « top-down » (MS-Align par exemple)(Liu, Sirotkin et al. 2011). Une recherche de PTM « à l'aveugle » peut ainsi être menée (Frank, Pesavento et al. 2008).

La spectrométrie de masse apporte aujourd'hui une multitude d'outils pour la recherche et la caractérisation des protéines glycosylées. En effet, grâce à l'approche « bottom-up » et aux spectromètres de masse de plus en plus performants, il est possible d'obtenir une liste exhaustive des protéines présentes dans un échantillon complexe. En adaptant l'approche « bottom-up » classique, avec utilisation éventuelle de méthodes d'enrichissement en glycopeptides ou glycoprotéines, et développement d'outils bioinformatiques dédiés, il semble aujourd'hui possible de déterminer la nature des glycoprotéines présentes au sein d'échantillons complexes afin de répondre à des questions biologiques d'intérêt.

De même, il semble également possible de caractériser de façon précise chacune de ces glycoprotéines, et notamment de déterminer la position de la glycosylation sur chacune d'entre elles. Pour cela, des techniques adaptées ont été développées, telles que l'analyse de protéines entière sur des appareils à très haute résolution, l'approche « top-down » de fragmentation des protéines, et le développement du mode de fragmentation ETD.

Ainsi, nous avons essayé d'évaluer l'apport de telles techniques pour la recherche et la caractérisation des glycoprotéines de *Mycobacterium tuberculosis*.

Introduction deuxième partie: L'application biologique

A. La tuberculose humaine

La tuberculose humaine, maladie infectieuse, est causée par la bactérie *Mycobacterium tuberculosis*. Elle se manifeste principalement sous forme d'infection pulmonaire mais d'autres formes minoritaires (osseuses, méningées, ganglionnaires) sont également observées.

Mycobacterium tuberculosis appartient au genre *Mycobacterium*, à la famille des *Mycobacteriaceae*, et à l'ordre des *actinomycetales*.

Il existe plus de 100 espèces au sein du genre *Mycobacterium* (Tortoli 2006) dont seulement 4 qui sont pathogènes : le complexe *tuberculosis*, *Mycobacterium leprae* responsable de la lèpre, *Mycobacterium ulcerans* responsable de l'ulcère de Buruli et *Mycobacterium marinum*. Il est intéressant de remarquer que toutes ces espèces sont des espèces à croissance lente.

Au sein du complexe *tuberculosis*, on retrouve *Mycobacterium tuberculosis*, *Mycobacterium africanum*, *Mycobacterium canettii* et enfin *Mycobacterium bovis*, responsable de la tuberculose bovine.

Au sein du genre *Mycobacterium* sont retrouvées des espèces à croissance rapide, comme *Mycobacterium smegmatis*. Parmi ces espèces à croissance rapide, certaines sont saprophytes, occasionnellement opportunistes si le sujet est atteint d'une immunodépression locale ou systémique, comme *Mycobacterium abscessus*.

Les mycobactéries sont des bactéries Gram positif en forme de batonnets d'environ 2 à 5 µm de long. Elles se distinguent des autres bactéries Gram positif par l'architecture unique de leur enveloppe (introduction 2eme partie).

I. L'état actuel de la tuberculose dans le monde

Il a été envisagé que *Mycobacterium tuberculosis*, *Mycobacterium africanum*, *Mycobacterium canettii* et *Mycobacterium bovis* découlent d'un ancêtre commun qui serait apparu 15 000 à 35 000 années auparavant (Daniel 2006).

C'est en 1882 que Robert Koch démontre l'origine bactérienne de la tuberculose et en identifie l'agent étiologique en isolant le bacille qui porte aujourd'hui son nom (Daniel 2005). Avec cette découverte, Koch reçoit le prix Nobel de Médecine en 1905.

Depuis, *Mycobacterium tuberculosis* n'a cessé d'évoluer et de contaminer le monde entier, en atteignant des proportions épidémiques en Europe et en Amérique du Nord au cours des XVIIIème et XIXème siècles (Almeida Da Silva and Palomino 2011).

Au cours des 200 dernières années, la tuberculose a fait plus d'un milliard de morts, ce qui en fait la maladie infectieuse la plus mortelle (figure 22) (Paulson 2013).

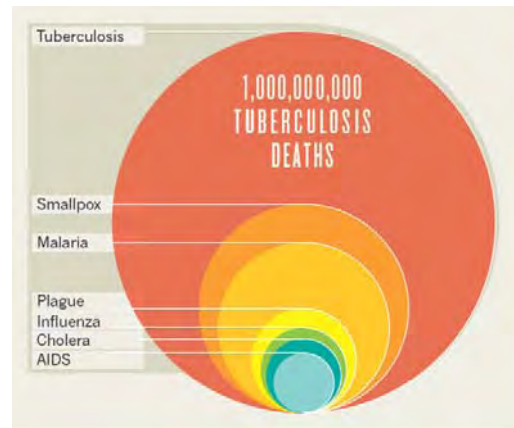


Figure 22: Les principales causes de mortalité par maladie infectieuse dans le monde au cours des deux cents dernières années. Figure de (Paulson 2013).

Chaque année, 9 millions de nouveaux cas de tuberculoses sont recensés, et près de 2 millions de morts. Près d'un tiers de la population mondiale est infectée par *Mycobacterium tuberculosis* et tous les pays sont touchés mais l'Afrique, l'Asie, l'Inde et la Chine sont les régions dans lesquelles l'on recense le plus de cas (OMS) (figure 23). De plus, la tuberculose est la première cause de décès chez les personnes atteintes du VIH.

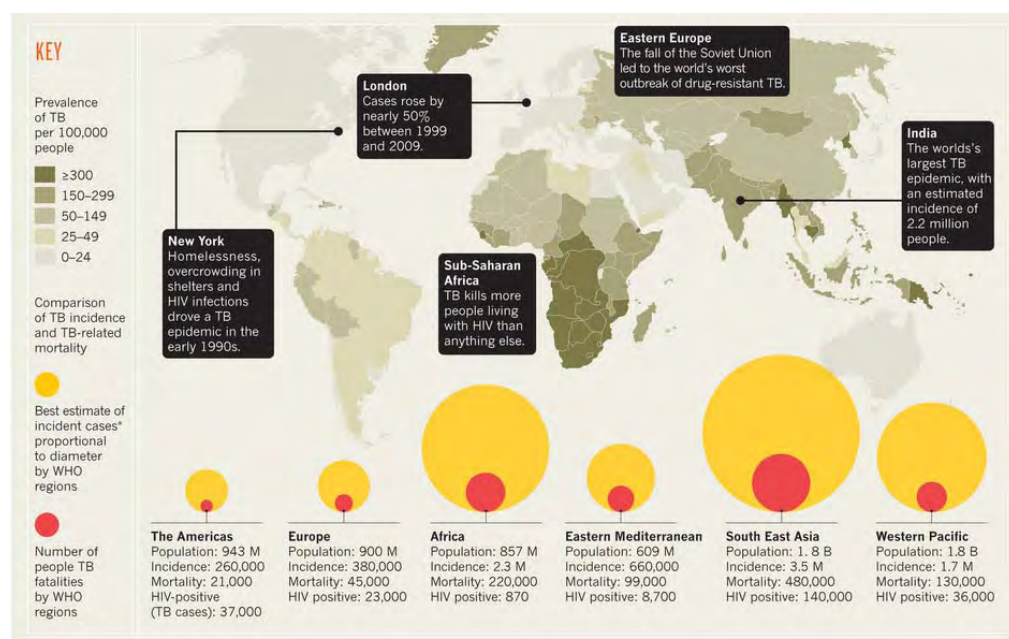


Figure 23: La répartition des victimes de la tuberculose selon les grandes régions du monde. Figure de (Paulson 2013).

Au cours des 15 dernières années, l'effort de la communauté mondiale pour lutter contre la tuberculose a permis une légère diminution du taux de mortalité due à cette maladie (figure 23). Afin de poursuivre sur cette voie, l'OMS a lancé un plan mondial « Halte à la tuberculose » sur 4 années, de 2011 à 2015.

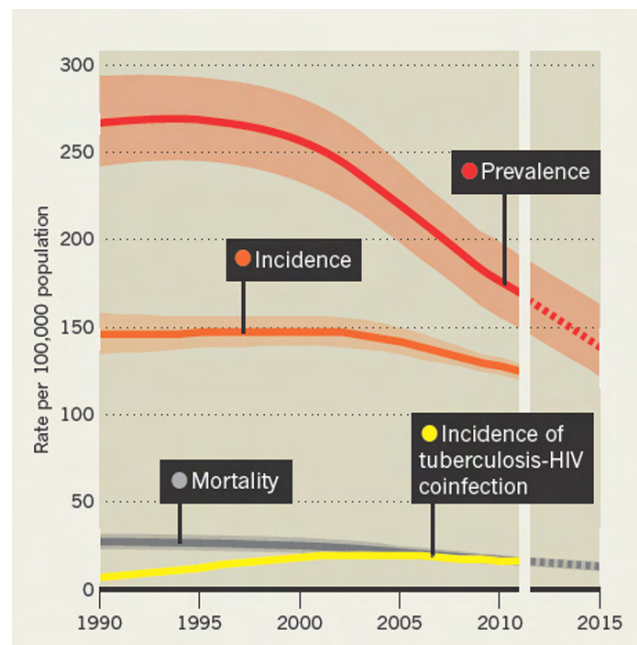


Figure 24: Le taux de mortalité dû à la tuberculose dans le monde a légèrement diminué au cours des quinze dernières années. Figure de (Paulson 2013).

II. Moyens de lutte

1. Le vaccin BCG

Un seul vaccin permettant une prévention contre la tuberculose a été développé à ce jour, c'est le vaccin BCG (Bacille de Calmette & Guérin). Il fut développé au début du XXème siècle par Albert Calmette et Camille Guérin.

Après 231 passages successifs de la souche *Mycobacterium bovis* sur un milieu de culture contenant de la glycérine et de la bile de bœuf, Camille Guérin et Albert Calmette obtiennent une souche atténuée dans sa virulence, incapable de produire des lésions tuberculeuses sur les animaux de laboratoire. C'est ensuite en 1921 qu'ils procèdent à la vaccination d'un enfant nouveau-né dans une famille de tuberculeux. L'enfant survit et ne présente aucun signe de la pathologie.

Le BCG a ensuite été distribué à travers le monde et maintenu pendant des décennies par des passages en série de différents stocks de bactéries (Ritz, Hanekom et al. 2008). Il existe maintenant plusieurs variants du vaccin, avec des différences génotypiques et phénotypiques (Brosch, Gordon et al. 2007). La question de l'efficacité de chacun de ces variants a donc été posée (Davids, Hanekom et al. 2006), et l'efficacité du BCG chez l'enfant a été avérée mais son efficacité chez l'adulte est remise en question (Andersen and Doherty 2005).

Des variations d'efficacité du BCG selon les populations vaccinées ont été observées. Une protection très limitée ou nulle de la population indienne après vaccination à par exemple été observée (Russell, Barry et al. 2010). Ces variations peuvent être attribuées à des différences d'ordre génétique, à différents modes de vie ou de régimes alimentaires, mais aussi à la possibilité dans certaines zones du monde d'être exposé aux mycobactéries environnementales non tuberculeuses avant d'être vacciné (Fine 1995, Bourzac 2014). En effet, certaines de ces mycobactéries possèdent des phénotypes communs avec *Mycobacterium tuberculosis* comme notamment l'expression de protéines fortement immunogènes tel que l'antigène Ag85.

De ce fait, de nouveaux vaccins sont à l'étude afin d'améliorer ou de remplacer le BCG par d'autres vaccins plus efficaces sur toutes les populations. Une première piste pour améliorer l'efficacité du vaccin BCG est de construire des souches recombinantes de *Mycobacterium bovis* BCG surexprimant des facteurs de virulence de *Mycobacterium tuberculosis*, ce qui permettrait d'amplifier la réponse immunitaire protectrice engendrée par l'administration du vaccin.

Une autre piste serait d'utiliser des souches de *Mycobacterium tuberculosis* atténuées dans leur virulence, après délétion de certains gènes de voies métaboliques requises pour la virulence ou la survie du bacille.

2. L'antibiothérapie

Le premier traitement de la tuberculose a débuté en 1946 avec l'utilisation de la streptomycine. Cette substance naturelle, un aminoglycoside, fut découverte en janvier 1944 par Selman Abraham Waksman. La streptomycine inhibe l'initiation de la traduction en se liant à l'ARN 16S de la sous-unité 30S des ribosomes mycobactériens. A cette époque, la streptomycine administrée seule a permis une réduction très importante de la mortalité immédiate mais cette monothérapie a engendré l'apparition de souches résistantes (Almeida Da Silva and Palomino 2011, Mitchison and Davies 2012). Dès 1948, la streptomycine a alors ensuite été administrée avec une autre drogue, l'acide para-amino-salicylique, ce qui a permis d'améliorer l'efficacité du traitement anti-tuberculeux.

A partir de 1952, l'isoniazide a également été introduite comme nouvelle drogue antituberculeuse (tableau 1). Cette substance altère la synthèse des acides mycoliques présents au niveau de la paroi bactérienne (figure 25). Son efficacité a été évaluée au cours d'une monothérapie ou au cours d'une trithérapie avec la streptomycine et l'acide para-amino-salicylique. Cette trithérapie, administrée au malade sur une durée d'un an, a été largement adoptée en Europe. La longue durée du traitement et son coût élevé n'ont cependant pas permis qu'il soit adopté dans les pays les plus pauvres.

Dans les années 1950, la capacité de la pyrazinamide à tuer des bacilles qui persistent au sein d'organes après traitement à l'isoniazide ou à la streptomycine est montrée. Le pyrazinamide induit une perturbation de la bactérie au niveau de sa membrane plasmique.

Un peu plus tard, en 1966 et 1972, il est montré que l'administration d'éthambutol et de rifampicine au cours d'une thérapie classique avec l'isoniazide et la streptomycine permet une réduction importante du taux de mortalité chez les patients. L'éthambutol est actif sur les bactéries en phase de croissance et altère la synthèse d'un des composants de la paroi, l'arabinogalactane. La rifampicine est efficace sur les bactéries en croissance et cible la sous-unité β de l'ARN polymérase.

Antibiotique	Cible	Mode d'action
Isoniazide	Enoyl-reductase	Inhibition de la synthèse des acides mycoliques
Pyrazinamide	Sous-unité 30S ribosomale	Inhibition de la synthèse protéique et acidification cytoplasme
Rifampicine	Sous-unité β ARN polymérase	Inhibition de la transcription
Ethambutol	Arabinosyl transférase	Inhibition de la synthèse d'arabinogalactane

Tableau 1: Les antibiotiques de première ligne utilisés pour le traitement contre la tuberculose

L'isoniazide, la pyrazinamide, l'éthambutol et la rifampicine constituent les quatre antibiotiques de première ligne du traitement de base de la tuberculose. Ce traitement dure 6 mois : 2 mois avec les quatre antibiotiques suivis de 4 mois avec l'isoniazide et la rifampicine.

Le taux de guérison des tuberculoses non multirésistantes est de 90% environ (Zumla, Nahid et al. 2013).

3. Souches MDR et XDR, et nouvelles drogues

Du fait de la longue durée du traitement antituberculeux de première ligne, un arrêt précoce de celui-ci par les patients survient souvent et a conduit à l'apparition de souches multi-résistantes aux antibiotiques (souches MDR, au moins résistantes à la rifampicine et à l'isoniazide).

Des antibiotiques de seconde intention sont utilisés pour le traitement de ces tuberculoses multi-résistantes : les fluoroquinolones (ciprofloxacine et ofloxacine par exemple), la kanamycine et l'amikacine qui sont des antibiotiques aminoglycoside, et la capréomycine et la vyomycine, des peptides cycliques. Ces antibiotiques sont plus toxiques que les antibiotiques de première intention et entraînent alors plus d'effets secondaires chez les patients. Des souches de *Mycobacterium tuberculosis* extrêmement résistantes aux traitements antibiotiques sont alors apparues (souches XDR : souches MDR résistantes aux antibiotiques de seconde ligne de type fluoroquinolones et à au moins trois drogues injectables dont la kanamycine par exemple).

Selon l'OMS, la tuberculose multi-résistante est présente dans tous les pays étudiés. 450 000 cas de tuberculoses multi-résistantes ont été recensés dans le monde en 2012, dont près de la moitié des cas en Inde, Chine et Russie. De plus, on estime que 9,6% des tuberculoses résistantes étaient des tuberculoses extrêmement résistantes. Les taux de guérison des tuberculoses multi-résistantes sont compris entre 50 et 70% seulement, contre 90% pour les tuberculoses non multi-résistantes.

La recherche s'oriente donc vers de nouveaux antibiotiques capables de lutter contre les souches MDR et XDR, et qui agissent sur la bactérie via plusieurs cibles pour accroître leur activité antibactérienne.

Parmi ces nouveaux antibiotiques, les nitroimidazoles ont montré une activité bactéricide sur des souches de *Mycobacterium tuberculosis* résistantes aux antibiotiques anti-tuberculeux classiques, comme les composés PA-824 (Stover, Warrenner et al. 2000) et OPC-67683 (Matsumoto, Hashizume et al. 2006) par exemple. Ils agissent sur les bactéries non répliquatives en état hypoxique en tant que donneur d'oxyde d'azote et causent aussi une chute des niveaux d'ATP. Le TMC207, lui aussi actif sur les souches MDR, agit comme un inhibiteur spécifique de l'ATP synthase mycobactérienne (Andries, Verhasselt et al. 2005). Une nouvelle classe d'agents antimycobactériens, les benzothiazinones (Makarov, Manina et al. 2009), éliminent *Mycobacterium tuberculosis* *in vitro* et *in vivo* chez la souris en ciblant la biosynthèse des polymères d'arabinane de l'enveloppe mycobactérienne.

B. *Mycobacterium tuberculosis* et sa pathogénicité

I. Structure de la bactérie

Les mycobactéries possèdent une enveloppe particulièrement imperméable de par sa teneur en lipides (environ 60% de la masse sèche). Cette enveloppe participe à la virulence des mycobactéries et à la résistance aux antibiotiques. Elle se compose de trois parties : une membrane plasmique, une paroi surmontée par une membrane externe, la mycomembrane, et une capsule (figure 25). Cette architecture de l'enveloppe a pu être mise en évidence par cryo-microscopie électronique. Hoffmann *et al.* (Hoffmann, Leis et al. 2008) et Zuber *et al.* (Zuber, Chami et al. 2008) ont ainsi pu démontrer pour la première fois l'existence d'une membrane externe ou mycomembrane, comparable à celle décrite chez les bactéries Gram négatif.

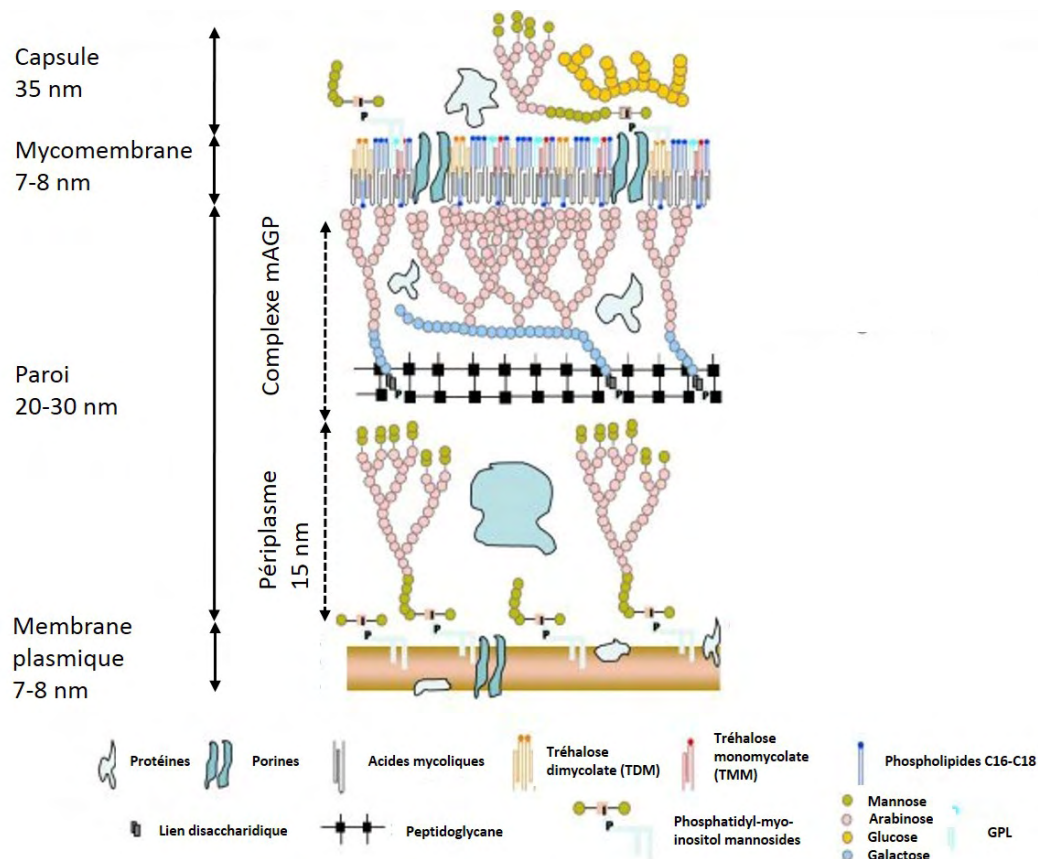


Figure 25: Les différents éléments composant l'enveloppe de *Mycobacterium tuberculosis*

1. La membrane plasmique

La membrane plasmique, couche la plus interne de l'enveloppe, est principalement constituée de phospholipides phosphatidylglycérol, phosphatidyléthanolamine,

phosphatidylinositol, estérifiés par des acides gras de 16 et 18 carbones et par un acide gras à 19 carbones, l'acide tuberculostéarique (C18 méthylé en position 10), spécifique des mycobactéries.

Les dérivés mannosylés du phosphatidylinositol, les phosphatidyl-*myo*-inositol mannosides (PIM) participent préférentiellement à la formation du feuillet externe de la membrane plasmique. Leur structure est basée sur une ancre phosphatidyl-*myo*-inositol mannoside (MPI) formée par un *sn*-glycérol-3-phospho-*myo*-inositol portant une unité α -D-mannopyranose en position 2 de l'inositol, et un domaine oligosaccharidique est greffé en position 6 de l'inositol (Chatterjee et al., 1992). Les formes majoritaires sont les PIM₂ et PIM₆. Quatre sites potentiels d'acylation des PIMs sont décrits et sont majoritairement estérifiés par des acides palmitique, stéarique et tuberculostéarique. L'ancre MPI permet l'ancrage des PIM dans la bicouche lipidique. Les PIM jouent un rôle important dans la perméabilité et l'intégrité de l'enveloppe mycobactérienne (Kordulakova, Gilleron et al. 2002, Morita, Sena et al. 2006) et leur synthèse est essentielle à la croissance des mycobactéries.

2. La paroi et le complexe mAGP

Les PIM sont les précurseurs de la biosynthèse du lipomannane (LM) et du lipoarabinomannane (LAM), lipoglycanes qui constituent le périplasme. Le LAM, dont le poids moléculaire est de 17 kDa chez *Mycobacterium bovis* BCG, est composé de trois domaines : une ancre MPI, un squelette polysaccharide et des coiffes. L'ancre MPI du LM et du LAM est identique à celle décrite pour les PIM.

Le peptidoglycane forme une couche rigide au-dessus du périplasme. Le peptidoglycane est composé de chaînes linéaires de N-acétyl-D-glucosamine et d'acide N-glycoyl-muramique. Le complexe covalent du peptidoglycane et de l'arabinogalactane mycoloylé (mAGP) constitue l'ossature de la paroi. L'arabinogalactane constitue le polysaccharide majeur de l'enveloppe mycobactérienne, il est constitué d'environ 125 résidus de D-galactofuranose et de D-arabinofuranose.

3. La mycomembrane

La mycomembrane est la couche la plus externe de la paroi, d'une épaisseur de 8 nm environ (Hoffmann, Leis et al. 2008).

Le feuillet interne de cette bicouche lipidique est formé par les acides mycoliques du complexe mAGP. Le feuillet externe est formé par des lipides extractibles, tels que des lipides à mycolates, et les lipoglycanes PIM, LM et LAM. Parmi les lipides extractibles, les

glycolipides à tréhalose sont constitués d'un cœur tréhalose estérifié par des acides gras, comme les tréhaloses mono- et dimycolates (TMM et TDM) et les di-, tri-, et polyacyltréhaloses (DAT, TAT et PAT respectivement) par exemple.

Les lipoglycanes LM et LAM, également présents au niveau de la membrane plasmique, sont également exposés à la surface des bactéries (Pitarque, Larrouy-Maumus et al. 2008) et sont probablement insérés dans la mycomembrane.

4. La capsule

La capsule constitue le compartiment le plus externe de l'enveloppe des mycobactéries. Son épaisseur est estimée à environ 35 nm. Elle est constituée majoritairement de protéines et polysaccharides (94-99%) et d'une faible quantité de lipides (1-6%) (Ortalo-Magne, Dupont et al. 1995). Les polysaccharides retrouvés dans la capsule sont l' α -D-glucane, l'arabinomannane et le mannane. L'arabinomannane et le mannane ont des structures similaires respectivement à celle du LAM et du LM, dépourvus d'ancrage MPI.

II. Physiopathologie des infections à *Mycobacterium tuberculosis*

Les poumons constituent la voie d'entrée du bacille tuberculeux chez l'hôte.

En effet, le bacille est inhalé sous forme de gouttelettes qui ont été libérées dans l'atmosphère après la toux d'une personne atteinte d'une tuberculose active. Ces gouttelettes sont d'une taille suffisamment petite pour rester suspendues dans l'atmosphère pendant quelques heures.

Le bacille atteint les alvéoles pulmonaires dans lesquelles il va être reconnu par le système immunitaire, et notamment par les macrophages alvéolaires (figure 25, 1). Ces derniers possèdent des récepteurs (PRR : Pattern Recognition Receptor) tels que les lectines de type C (récepteur au mannose, Mincle Dectin-1 et DC-SIGN), les récepteurs Toll-like (TLR) ou les récepteurs NOD capables de détecter la présence d'agents infectieux (voir partie II.C.). Ceci engendre alors une inflammation localisée et la sécrétion de cytokines par les macrophages infectés, qui conduit au recrutement de cellules mononucléaires (monocytes, cellules dendritiques par exemple) provenant des vaisseaux sanguins voisins (figure 25, 2). On observe également une néo vascularisation importante.

Les macrophages infectés sont alors entourés par d'autres macrophages dont certains se différencient en « foamy » macrophages riches en lipides, en macrophages epithélioïdes, ou en macrophages multinucléés, par des monocytes et des neutrophiles. Il y a formation du granulome qui constitue la preuve pathologique de la maladie et qui reste à ce stade une

masse amorphe. Des lymphocytes arrivent également au lieu de l'inflammation (figure 25, 3).

Avec l'arrivée des lymphocytes, le granulome acquiert une structure plus organisée avec un manteau fibreux qui se forme autour des macrophages et qui marque la périphérie du granulome (figure 25, 4). Au sein de ce granulome, les bactéries peuvent se multiplier à l'état extracellulaire.

Une fois la phase de réplication bactérienne rapide terminée, des foyers caséux apparaissent. Le bacille persiste dans ces foyers mais ne se réplique plus et on parle alors de tuberculose latente. Cet état correspond à un équilibre dynamique entre le système immunitaire de l'hôte et la bactérie qui peut perdurer tout au long de la vie de l'individu atteint. C'est le cas pour 90% des personnes ayant été infectées par *Mycobacterium tuberculosis*.

Suite à une diminution des défenses immunitaires de l'individu, cet état peut basculer au profit du pathogène, comme cela est le cas pour 10% des personnes infectées. La vascularisation du granulome diminue, le caseum s'accumule au centre du granulome. (figure 25, 5).

La rupture du granulome engendre la libération des bactéries au niveau des voies respiratoires, le bacille peut alors sortir de l'organisme par expectoration et infecter un autre individu (figure 25, 6).

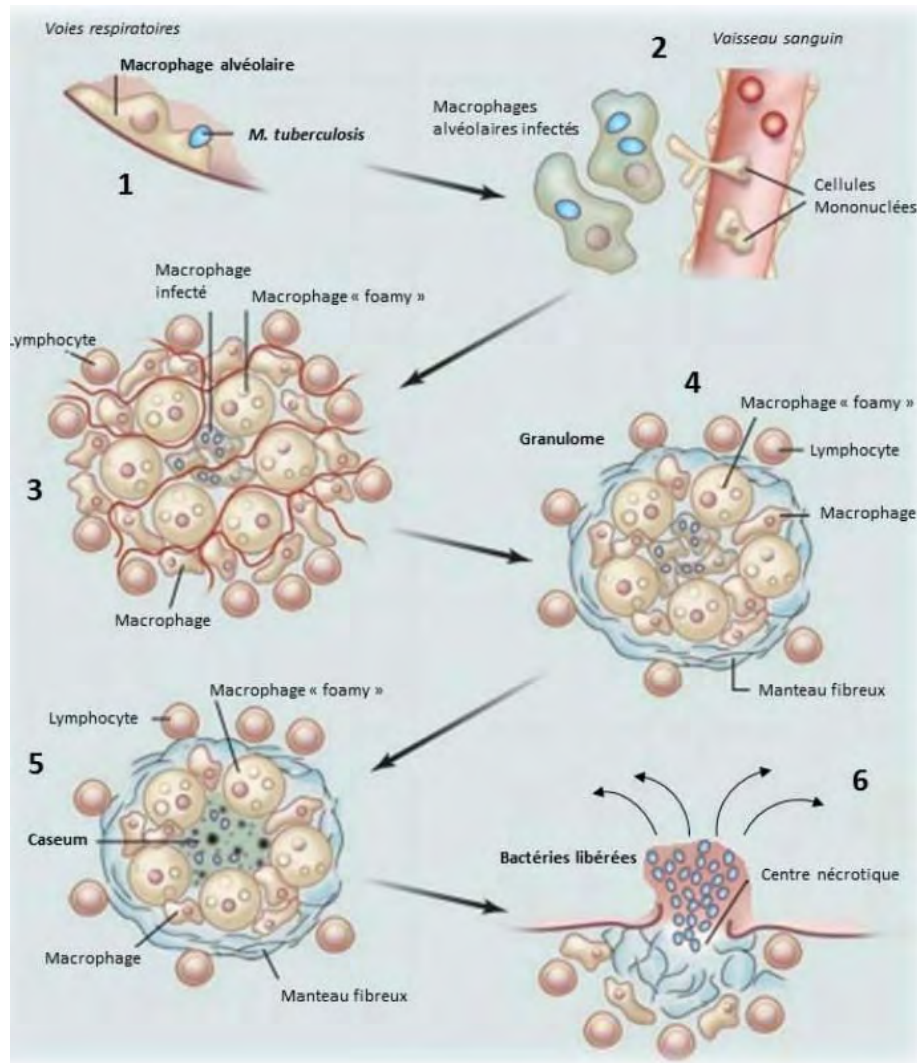


Figure 26: Les différentes étapes du cycle infectieux de *Mycobacterium tuberculosis* Figure de (Russell, Barry et al. 2010)

III. Moyens de reconnaissance de *Mycobacterium tuberculosis* par le système immunitaire inné

Le système immunitaire inné est la première ligne de défense contre les microorganismes. Contrairement au système immunitaire adaptatif, il est immédiatement actif et n'a pas besoin d'avoir rencontré l'agent infectieux au préalable. Une grande variété de cellules participe à la réponse immunitaire innée, notamment les cellules phagocytaires telles que les macrophages, les cellules dendritiques, les monocytes et les neutrophiles, mais aussi des cellules « Natural Killer » (NK), les éosinophiles, les basophiles et les cellules épithéliales. Ces cellules sont capables de détecter la présence des microbes, et de les internaliser pour induire une réponse inflammatoire et recruter les autres partenaires du système immunitaire au site de l'infection. La reconnaissance des microorganismes par le système immunitaire est basée sur la détection de motifs moléculaires associés aux pathogènes

(pathogen-associated molecular patterns, PAMP) par des récepteurs, les PRR présents à la surface des cellules immunitaires (Janeway and Medzhitov 2002, Medzhitov 2007).

1. Les récepteurs Toll-like

Les récepteurs Toll-like (TLR) sont des protéines trans-membranaires et constituent une famille de PRR, conservée chez tous les mammifères. A ce jour, 10 TLR ont été décrits chez l'homme et sont exprimés à la membrane plasmique (TLR1, 2, 4, 5 et 6) ou au niveau des compartiments endosomaux (TLR 3, 7, 8 et 9) des cellules immunitaires, notamment des macrophages et des cellules dendritiques.

Les protéines TLR sont constituées d'un domaine cytoplasmique TIR, un domaine transmembranaire et un domaine extracellulaire composé de motifs riches en leucines. Ce dernier permet la reconnaissance des ligands.

Les TLR de type TLR2, TLR4 et TLR9 peuvent interagir avec *Mycobacterium tuberculosis*. Cette interaction induit la sécrétion de cytokines dépendantes de TLR2, TLR4 et TLR9, telles que les cytokines pro-inflammatoires IL-12 et TNF α , ou de la cytokine anti-inflammatoire IL-10 (Carmona, Cruz et al. 2013).

Des études épidémiologiques se sont intéressées à l'association de facteurs génétiques de l'hôte avec la sensibilité ou la résistance à la tuberculose. Plusieurs polymorphismes de TLR2 ont été associés à une susceptibilité accrue pour la tuberculose. D'autres polymorphismes décrits pour TLR1, TLR6 et TLR4 ont également été associés à une augmentation des cas de tuberculose, soulignant le possible rôle des TLR lors d'une infection par *Mycobacterium tuberculosis* (Kleinnijenhuis, Oosting et al 2011).

2. Les lectines de type C

Les lectines de type C sont des protéines capables de lier les sucres de manière dépendante des ions divalents Ca²⁺. La liaison aux sucres est localisée au niveau du domaine de reconnaissance des sucres (carbohydrate-recognition domain, CRD) dont les structures secondaires et tertiaires sont partagées par toutes les lectines. Le récepteur au mannose, DC-SIGN (Dendritic cell-specific ICAM-3-grabbing non integrin), Mincle (macrophage indicible C-type lectin) et Dectin-1, des lectines du type C exprimées à la surface des cellules du système immunitaire sont impliqués dans la reconnaissance des pathogènes. De façon générale, l'activation des récepteurs lectiniques aboutit à l'endocytose, au trafic intracellulaire et à la transduction du signal d'infection lors des premières étapes d'une infection.

- Le récepteur au mannose

Le récepteur au mannose possède huit domaines CRD. Ce récepteur reconnaît les composés glycosylés de l'enveloppe mycobactérienne tels que le ManLAM (Schlesinger, Hull et al. 1994), l'arabinomannane, le mannane, les PIM (Torrelles, Azad et al. 2006) et les protéines mannosylées. L'interaction de *Mycobacterium tuberculosis* avec le récepteur au mannose lui permet son entrée dans le macrophage (Ernst 1998). De plus, l'interaction du lipoarabinomannane à coiffes mannosides (ManLAM) de *Mycobacterium bovis BCG* ou de *Mycobacterium tuberculosis* avec le récepteur au mannose induit l'inhibition de la sécrétion de la cytokine pro-inflammatoire IL-12 par des cellules dendritiques stimulées au LPS (Nigou, Zelle-Rieser, 2001) et l'inhibition de la fusion du phagosome contenant la mycobactérie avec le lysosome (Kang, Azad et al. 2005), ce qui montre le rôle du récepteur au mannose dans la modulation de la réponse immunitaire.

- DC-SIGN

DC-SIGN est le récepteur majeur de *Mycobacterium tuberculosis* et *Mycobacterium bovis BCG* à la surface des cellules dendritiques et reconnaît spécifiquement les mycobactéries du complexe *tuberculosis*. DC-SIGN possède un seul CDR.

L'interaction du ManLAM avec DC-SIGN permet de moduler la réponse des cellules dendritiques dont la maturation et l'expression des marqueurs de surface CD83 et CD86 sont diminuées. Cette interaction inhibe également la sécrétion de la cytokine pro-inflammatoire IL-12 et induit la sécrétion de la cytokine anti-inflammatoire IL-10 (Geijtenbeek, Van Vliet et al. 2003).

- Mincle et Dectin-1

Mincle est exprimé à la surface des macrophages (Matsumoto, Tanaka et al. 1999), et interagit avec le théalose dimycolate, glycolipide de la membrane externe de l'enveloppe mycobactérienne. Cette interaction conduit à l'activation des macrophages avec la production de NO et de TNF.

Dectin-1 est connue pour son rôle de détection des β -glucanes de pathogène fongiques, mais aucun polysaccharide mycobactérien possédant cette structure n'est connu. Jusqu'à présent, aucun ligand mycobactérien de Dectin-1 n'a été identifié.

C. La O-mannosylation des protéines mycobactériennes

I. Le mécanisme de O-glycosylation

La preuve de la glycosylation des protéines chez les procaryotes a été donnée en 1976 avec la première caractérisation de la glycosylation d'une protéine de surface de la bactérie gram-négative *Halobacterium salinarium* (Mescher 1976).

Seules des *N*-glycosylations et des *O*-glycosylations ont été observées chez les procaryotes, alors que des *S*-glycosylations et *C*-glycosylations existent également chez les eucaryotes. Des motifs saccharidiques communs bien caractérisés chez les eucaryotes (glucose, galactose, N-acétylgalactosamine, etc...) peuvent également être observés chez les procaryotes, mais d'autres motifs plus rares sont également observés (acide legionaminique par exemple) (Abu-Qarn, Eichler et al. 2008).

Chez *Mycobacterium tuberculosis*, seules des protéines *O*-mannosylées ont à ce jour été décrites. En 1989, Espitia et Mancilla (Espitia and Mancilla 1989) mettent en évidence trois protéines de *Mycobacterium tuberculosis* ayant une affinité pour la Concanavaleine A, et composées à la fois d'acides aminés et d'unités osidiques. La première identification de protéines glycosylées de *Mycobacterium tuberculosis* est alors faite.

En 1996, Dobos *et al.* caractérisent la glycosylation de la protéine Apa (Alanine and Proline rich Antigen), antigène sécrété de la bactérie. Les peptides glycosylés sont identifiés avec quatre thréonines comme sites de glycosylation (Thr49, Thr57, Thr66 et Thr316), porteuses de di-mannoses ou tri-mannoses, avec des liaisons de type $\alpha 1,2$. Cette étude représente la première caractérisation d'une glycoprotéine mycobactérienne (Dobos, Khoo et al. 1996).

En 1996 également, Hermann *et al.*, via une mutagénèse dirigée (sur les thréonines Thr34, Thr35, Thr36, Thr40 et 41) et l'affinité de la protéine pour la Concanavaleine A mettent en évidence la glycosylation de la lipoprotéine LpqH (Herrmann, O'Gaora et al. 1996).

En 2003, la protéine MPB83 de *Mycobacterium bovis* est également identifiée comme protéine glycosylée, avec 3 mannoses répartis sur deux thréonines, liés par des liaisons de type $\alpha 1,3$ (Michell, Whelan et al. 2003).

En 2009, la glycosylation de la protéine superoxide dismutase SodC est caractérisée et six acides aminés modifiés sont identifiés avec six sites de glycosylation (Thr45, thr46, Ser48, Thr51, Ser53 et ser56), porteurs de 1 à 3 mannoses avec des liaisons de type $\alpha 1,2$, et un maximum de 9 mannoses observés sur la protéine (Sartain and Belisle 2009).

C'est en 2009 également que Gonzalez-Zamorano *et al.* réalisent un gel d'électrophorèse bi-dimensionnel avec les protéines du surnageant de culture du bacille et évaluent l'affinité de ces protéines pour la concanavaline A. Ce sont alors 41 protéines qui possèdent une affinité pour la lectine et qui sont identifiées par spectrométrie de masse. En revanche, les peptides glycosylés ne sont pas identifiés au cours de cette étude.

Du fait de la difficulté inhérente à la recherche et à l'analyse de protéines glycosylées, peu de glycoprotéines mycobactériennes étaient alors décrites et totalement caractérisées au début de ces travaux de thèse.

Chez les eucaryotes et les procaryotes, la *O*-mannosylation est initiée par le transfert d'une première unité mannoside d'un lipide donneur de mannose vers une thréonine ou une sérine.

Chez les eucaryotes, une unité mannoside est transférée d'un donneur de mannose GDP-Man vers une molécule de dolichol-phosphate ancrée dans la membrane de l'appareil de Golgi, pour former un dolichol-phosphate mannose (Dol-P-man). Il y a ensuite translocation de ce Dol-P-man vers le cytoplasme. Le transfert de l'unité mannoside du Dol-P-man vers la première thréonine ou sérine de la glycoprotéine naissante est ensuite catalysé par une protéine *O*-mannosyl-transférase, intégralement membranaire. Sept protéines *O*-mannosyl-transférases ou PMTs ont été identifiées chez *Saccharomyces cerevisiae* (Strahl Bolsinher 1999). L'élongation des chaînes saccharidiques se fait ensuite à l'aide d' α 1,2 ou α 1,3 mannosyltransférases, de la famille des KTR par exemple (ScKre2p, ScKtr1p, ScKtr1p, α 1,2 mannosyltransférases). Le GDP-Man joue le rôle de donneur de mannose pour cette réaction d'élongation de la chaîne saccharidique.

En 2005, VanderVen *et al.*, par une comparaison bio-informatique des génomes de *Mycobacterium tuberculosis* et de *Saccharomyces cerevisiae* ont pu identifier une unique protéine de *Mycobacterium tuberculosis* ayant un profil d'hydropathie similaire aux PMTs de *Saccharomyces cerevisiae* et 24% d'identité avec celles-ci. Cette protéine est codée par le gène Rv1002c. Il est à noter de façon intéressante que le gène Rv1002c est conservé dans le génome de toutes les mycobactéries et en particulier de celle de *Mycobacterium leprae*. Or, il est considéré que cette souche mycobactérienne possède le génome minimum requis pour sa survie (Vissa and Brennan 2001), ce qui suggère un rôle important du gène Rv1002c dans la survie des mycobactéries.

En revanche, au début de ces travaux de thèse, rien n'était encore connu sur la ou les protéine(s) de *Mycobacterium tuberculosis* responsable(s) de l'élongation des chaînes saccharidiques.

De plus, il est également montré dans cette étude de VanderVen *et al.* que la *O*-mannosylation des protéines chez *Mycobacterium tuberculosis* est couplée au système Sec

d'export des protéines, ce qui suggère que seules des protéines exportées du bacille peuvent être *O*-mannosylées.

Enfin, il n'existe pas de séquence consensus pour prédire les sites de *O*-glycosylation. En revanche, il a été montré que des sérines et thréonines avec un environnement riche en prolines ont une plus grande probabilité d'être *O*-glycosylées (Gooley, Classon et al. 1991)

II. Le rôle de la glycosylation des protéines chez *Mycobacterium tuberculosis*

Le rôle de la mannosylation des protéines chez les eucaryotes est bien documenté. Chez l'humain, une déficience en la synthèse de motifs *O*-mannosidiques a de graves conséquences et entraîne des formes sévères de dystrophies musculaires congénitales.

Chez la levure, et notamment chez *Saccharomyces cerevisiae*, la *O*-mannosylation des protéines est essentielle pour sa viabilité, et en particulier pour l'intégrité de la membrane plasmique. Chez *Candida albicans*, espèce pathogène, une diminution du taux de glycosylation des protéines entraîne une diminution de l'adhérence aux cellules hôtes, et une diminution de la virulence chez la souris.

En revanche, peu de choses sont connues sur le rôle fonctionnel de la mannosylation des protéines chez les procaryotes. Chez *Corynebacterium glutamicum* et *Streptomyces coelicolor*, actinomycètes proches des mycobactéries dans l'arbre phylogénétique, les mutants dépourvus du processus de mannosylation des protéines sont viables et ne montrent aucun phénotype particulier (Mahne, Tauch et al. 2006, Wehmeier, Varghese et al. 2009).

En ce qui concerne les mycobactéries, dans une étude de 1993, la protéine Apa est isolée au cours d'une recherche des antigènes majeurs de *Mycobacterium bovis* BCG. Cette protéine induit une forte réaction d'hypersensibilité retardée chez des cochons d'Inde, 24 à 48 heures après immunisation avec le vaccin BCG. Il est alors montré que la protéine Apa est capable de stimuler les lymphocytes T.

En 1999, il est montré que la capacité de la protéine Apa à induire une réponse des lymphocytes T CD4+ et CD 8+ *in vitro* et *in vivo* est dépendante de son taux de glycosylation. En effet, la protéine non-glycosylée produite chez *Escherichia coli* est incapable de produire une réaction d'hypersensibilité retardée (Horn, Namane et al. 1999, Romain, Horn et al. 1999). Il semblerait alors qu'une diminution du taux de glycosylation de Apa diminue sa capacité à entrer dans les cellules hôtes et à jouer son rôle protectif.

De plus, il a également été montré au laboratoire que les motifs glycosidiques de la protéine Apa lui permettent de se lier à la lectine de type-C du système immunitaire PSP-A. Les motifs

de la protéine Apa jouent alors un rôle déterminant dans la colonisation et l'invasion des cellules de l'hôte par le pathogène (Ragas, Roussel et al. 2007).

En 2006, Sartain *et al.* identifient SodC comme un antigène présent dans le surnageant de culture de *Mycobacterium tuberculosis* pouvant être utilisé dans le cadre d'un sérodiagnostic de la maladie (Sartain, Slayden et al. 2006).

En 2008, la glycosylation de la protéine LprG de *Mycobacterium leprae* a été montrée comme indispensable pour l'activation de la classe de lymphocytes T MHC II chez des patients atteints de la lèpre (Sieling, Hill et al. 2008).

Enfin, la glycosylation des protéines de *Mycobacterium tuberculosis* (Herrmann, 1996) pourrait être un moyen de protection face à la protéolyse. En effet, il est montré que la protéine non glycosylée (produite chez *Escherichia coli*) est plus sensible à la protéolyse que la protéine glycosylée.

Les glycoprotéines semblent alors être des éléments clés de la bactérie lui permettant d'infecter les cellules hôtes via des interactions avec les lectines de type C. Cependant, aucune des glycoprotéines identifiées à ce jour n'a été montrée comme indispensable pour la virulence du bacille tuberculeux. De plus, aucune donnée sur l'effet systémique de la glycosylation des protéines n'a été donnée.

Même si le processus de glycosylation des protéines a été montré comme non indispensable chez les actinomycètes, le processus de glycosylation de *Mycobacterium tuberculosis* possède des similarités avec celui des eucaryotes, ce qui peut nous amener à penser qu'un tel processus, qui requiert une consommation d'énergie non négligeable, est requis par la bactérie pour se maintenir au sein de l'hôte.

Le blocage du processus de O-mannosylation apparaît aujourd'hui comme une cible thérapeutique intéressante contre les bactéries et champignons pathogènes encore très peu explorée (Lengeler, Tielker et al. 2007).

Le but de ce travail est alors d'identifier d'autres glycoprotéines de la bactérie pouvant jouer un rôle important dans la virulence du bacille. Au cours de la thèse, en 2013, est parue l'étude de Smith *et al.* qui constitue la première approche globale d'identification des glycoprotéines sécrétées du bacille et des glycopeptides associés. 13 glycoprotéines ont ainsi été identifiées, en combinant diverses approches de spectrométrie de masse. De la même façon, nous allons utiliser une approche de glycoprotéomique faisant intervenir un outil bioinformatique développé au laboratoire afin d'identifier les glycopeptides du bacille, et nous utiliserons des stratégies de spectrométrie de masse innovantes pour une caractérisation complète de glycoprotéines mycobactériennes.

PRESENTATION GENERALE DES TRAVAUX

Nous avons vu dans l'introduction le rôle important de la *O*-mannosylation des protéines bactériennes dans la virulence du bacille tuberculeux. Le but de ma thèse a donc été de mettre en place des outils protéomique et de spectrométrie de masse pour l'identification et la caractérisation des mannoprotéines qui contribuent à la virulence de *M. tuberculosis*.

Pour cela, au cours de cette thèse, je me suis attachée à développer une méthode efficace d'identification des glycoprotéines au sein de mélanges complexes basée sur une approche originale de retraitement des données de spectrométrie de masse. Nous verrons donc dans une première partie de ce manuscrit l'utilisation du programme bioinformatique « Neutral Loss » qui consiste à rechercher, parmi tous les spectres MS/MS acquis au cours d'une stratégie « bottom-up » classique, ceux qui possèdent une signature de peptides glycosylés. J'ai optimisé l'utilisation de cet outil tout au long de la thèse afin d'obtenir une identification des glycoprotéines fiable. Durant cette phase de développement nous avons évalué l'intérêt de cet outil en le confrontant à deux méthodes d'enrichissement en glycopeptides ou glycoprotéines plus conventionnellement utilisées pour la recherche et l'identification de glycoprotéines.

L'utilisation de cet outil bioinformatique « Neutral Loss » nous a permis de rechercher les glycoprotéines de la bactérie sécrétées dans deux milieux de culture différents, mais également d'analyser les glycoprotéines associées aux corps bactériens et obtenues après lyse des bactéries. Enfin, nous avons également appliqué la même stratégie de recherche des glycoprotéines chez une souche mutante du bacille $\Delta Rv1002c$ délétée du gène codant pour la Protein Mannosyl Transferase, responsable de la mannosylation des protéines, afin de contrôler et valider notre processus analytique.

Par ailleurs, pour comprendre le rôle global de la mannosylation des protéines chez la bactérie, nous avons également mis en place une étude comparative et quantitative des protéines de la souche sauvage et de la souche mutante $\Delta Rv1002c$ (ainsi que de la souche complémentée correspondante) afin de rechercher des différences d'expressions de protéines dans ces deux souches pouvant conduire à la mise en évidence de voies métaboliques affectées.

Pour finir, nous verrons également les stratégies d'analyse complète de glycoprotéines que nous avons mis en place afin de pouvoir caractériser finement les glycoprotéines d'intérêt dégagées au cours de l'étude. Pour cela, nous avons travaillé avec une protéine modèle, la Fasciline de *Mycobacterium smegmatis* sur laquelle nous avons évalué l'apport des méthodes « bottom-up » et « top-down » avec fragmentation ETD pour la localisation

des sites de glycosylation, mais aussi la méthode de spectrométrie de masse à mobilité ionique pour la séparation de glycoformes isobares.

RESULTATS Première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

A. Le contexte biologique

Afin d'évaluer l'effet d'une absence de glycosylation des protéines sur la viabilité et la virulence de *Mycobacterium tuberculosis*, la souche mutante $\Delta Rv1002c$ (gène codant pour la PMT délété) a été construite en collaboration avec l'équipe de C. Guilhot à l'IPBS. De plus, la souche mutante $\Delta Msmeg_5447$ de la souche saprophyte *Mycobacterium smegmatis*, délétée du gène codant pour la PMT homologue a également été construite, ainsi que les souches complémentées correspondantes des deux espèces. Les phénotypes de croissance de ces deux souches mutantes ont été caractérisés, ainsi que le phénotype de virulence de la souche mutante $\Delta Rv1002c$.

I. Impact de la mannosylation des protéines sur la viabilité et la croissance des souches $\Delta Msmeg_5447$ et $\Delta Rv1002c$

La croissance du mutant $\Delta Msmeg_5447$ en milieu de culture 7H9 avec dextrose et catalase (7H9 DC) est similaire à celle de la souche sauvage et ne semble donc pas être impactée par un arrêt du processus de glycosylation (figure 27A).

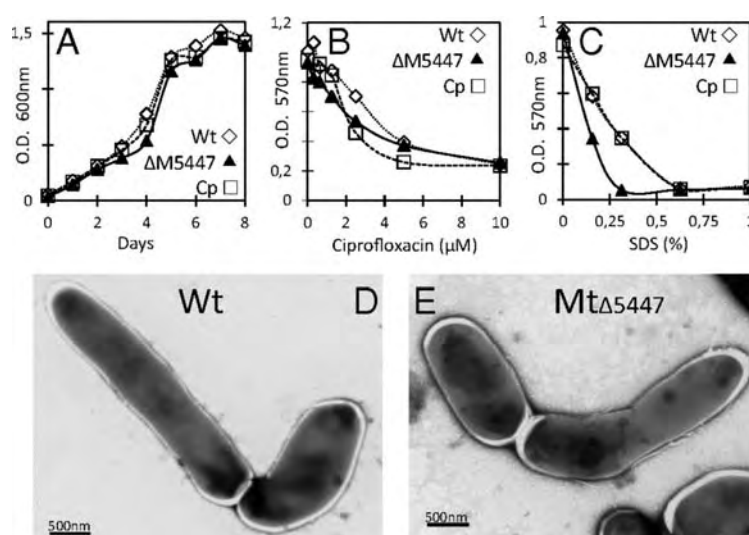


Figure 27: Effets de la délétion du gène *Msmeg_5447* sur la croissance de la bactérie (A), la susceptibilité à la ciprofloxacine (B), et au stress induit par l'agent chaotrope SDS (C). Observation des bactéries de la souche sauvage (D) et de la souche mutante (E) au microscope électronique.

De plus, la susceptibilité aux antibiotiques de cette souche mutante est semblable à celle de la souche sauvage, ce qui traduit le fait que la délétion du gène *Msmeg_5447* n'affecte pas la perméabilité de la membrane aux drogues (figure 27B). Enfin, la tolérance aux agents chaotropiques de la souche mutante diminue par rapport à celle de la souche sauvage, ce qui pourrait être en faveur d'un changement de l'architecture de l'enveloppe (figure 27C). Cependant, l'observation des deux bactéries sauvage et mutante au microscope électronique ne permet de détecter aucun changement au niveau de l'enveloppe.

En revanche, la croissance du mutant $\Delta Rv1002c$ de *Mycobacterium tuberculosis* est fortement impactée par l'arrêt de l'expression de la PMT (figure 28). En effet, dans un milieu de culture liquide 7H9 avec dextrose et catalase (7H9 DC), la croissance de la souche mutante diminue fortement par rapport à celle de la souche sauvage. Cependant, dans un milieu de culture liquide 7H9 albumine, dextrose et catalase (7H9 ADC), la croissance de la souche mutante est comparable à celle de la souche sauvage.

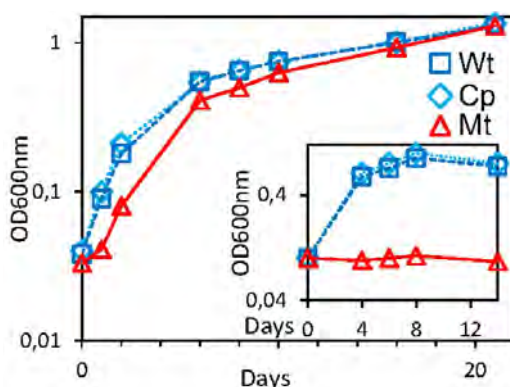


Figure 28: Impact de la délétion du gène $\Delta Rv1002c$ sur la croissance de *Mycobacterium tuberculosis*, dans un milieu 7H9 ADC et 7H9 DC (insertion)

II. Impact de la mannosylation des protéines sur la virulence de la souche *Rv1002c*

Les capacités de la souche sauvage et de la souche mutante $\Delta Rv1002c$ de *Mycobacterium tuberculosis* à entrer et se maintenir dans les macrophages ont été comparées.

Des macrophages alvéolaires murins ainsi que des macrophages humains dérivés de monocytes ont été infectés (une bactérie par macrophage). En ce qui concerne les macrophages murins, la vitesse initiale d'infection est significativement plus faible pour la souche mutante que pour la souche sauvage (figure 29 A et B). Cette observation semble cohérente avec le fait que des glycoprotéines exposées à la surface, comme la protéine Apa, peuvent jouer le rôle d'adhésines sur la membrane des macrophages et interagir avec les lectines de type C.

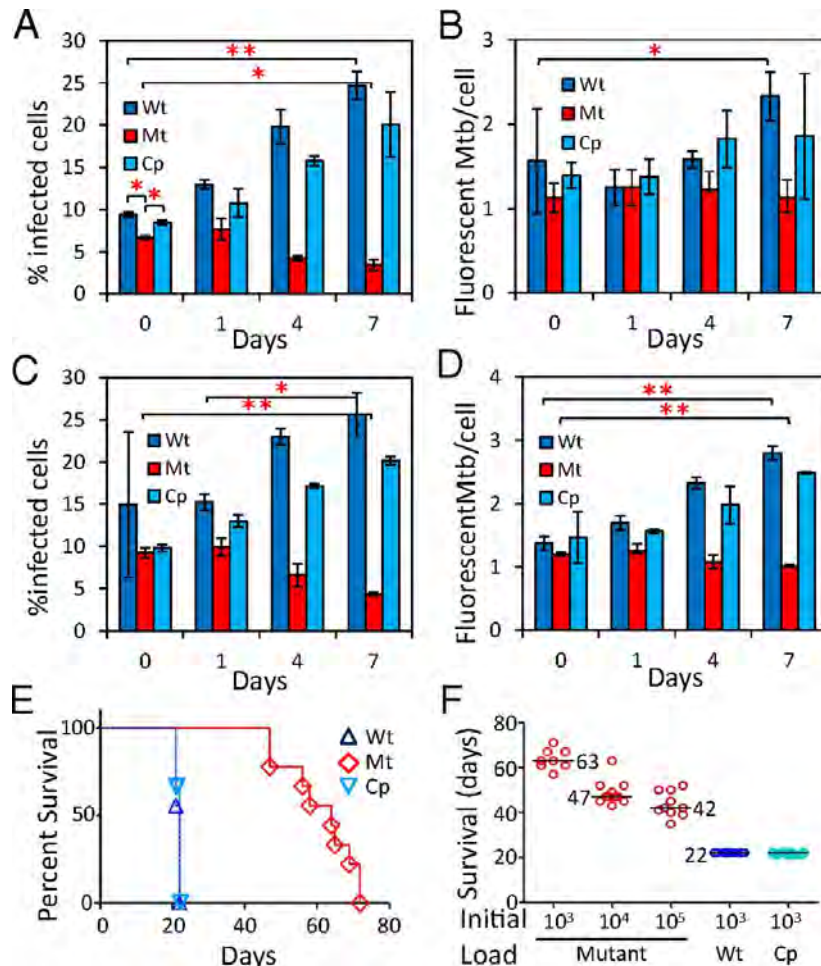


Figure 29: Dans le cas de la souche mutante, le nombre de bactéries par macrophage murin ou humain n'augmente pas au cours du temps après infection (B murin et D humain), et le pourcentage de cellules infectées diminue (A murin et C humain). Ces observations semblent en faveur d'une croissance intracellulaire de la souche mutante limitée. Chez la souris SCID immunodéprimée, la virulence de la souche mutante est fortement diminuée par rapport aux souches sauvage et complétée: les souris SCID commencent à mourir à partir de 40 jours lorsqu'elles sont infectées avec la souche mutante contre 20 jours avec la souche sauvage ou la souche complétée (E). De même, la médiane de survie des souris est plus élevée après infection avec la souche mutante qu'après infection avec la souche sauvage ou la souche complétée (F).

Afin d'évaluer la virulence de la souche mutante *in vivo*, des souris SCID immunodéprimées ont été infectées par voie intranasale, avec des bactéries provenant de la souche sauvage, de la souche mutante ou de la souche complétée.

Les souris infectées avec la souche sauvage ne survivent pas plus de 22 jours alors qu'aucune mort des souris infectées avec la souche mutante n'est observée avant 47 jours (figure 29 F). D'autres infections ont été réalisées avec des doses croissantes de bactéries et la moyenne de survie a été montrée comme dépendante de cette charge initiale.

En extrapolant les médianes de survie en fonction de la charge bactérienne initiale, la virulence de la souche mutante chez des souris SCID immunodéprimées est alors montrée comme diminuée d'un facteur 10^3 à 10^4 par rapport à la virulence de la souche sauvage.

Le mutant Δ Rv1002c est donc sévèrement impacté par un arrêt du processus de glycosylation. Pour essayer de comprendre ce phénomène, nous avons entrepris de rechercher d'autres glycoprotéines que celles précédemment identifiées dans la littérature qui pourraient expliquer ce phénotype. En effet, les mutants de chacune de ces glycoprotéines connues ont été construits et aucun ne possède le phénotype observé.

B. Identification des glycoprotéines sécrétées de *Mycobacterium tuberculosis* – Milieu de culture 7H9

I. Echantillons biologiques

Comme il a précédemment été évoqué (introduction deuxième partie, partie C.I), la protéine PMT a été montrée comme couplée au système d'export SEC des protéines. Nous pouvons alors nous attendre à ce que seules des protéines extra-cytoplasmiques soient potentiellement glycosylées. Les glycoprotéines du bacille tuberculeux ayant été caractérisées avant ce travail de thèse ((introduction deuxième partie, partie C.I., les protéines Apa, SodC, MPB83, LpqH) sont des protéines retrouvées dans le surnageant de culture de la bactérie.

Nous avons donc choisi dans un premier temps de rechercher les glycoprotéines dans le surnageant de culture provenant des milieux 7H9 ADC et 7H9 DC, avec un temps de culture de deux semaines.

En effet, avant le début de ce travail de thèse, des cultures des trois souches de *Mycobacterium tuberculosis* en milieu 7H9 ADC et 7H9 DC avaient été lancées. De façon générale, les conditions de culture dans ce milieu sont bien maîtrisées. Cependant, il n'a pas été possible de cultiver la souche mutante dans un milieu de culture 7H9 DC. Or, il était nécessaire de pouvoir cultiver la souche mutante afin de pouvoir analyser les protéines sécrétées par cette souche, et contrôler et valider notre processus analytique de recherche des glycoprotéines. Cependant, la présence d'albumine en excès dans le milieu de culture 7H9 ADC n'est pas confortable pour une analyse protéomique.

Tout en testant d'autres milieux de culture permettant de cultiver les 3 souches simultanément, il a été décidé de commencer par analyser le glycoprotéome de *Mycobacterium tuberculosis* dans les deux milieux 7H9 DC et 7H9 ADC pour évaluer un éventuel rôle de l'albumine dans la glycosylation des protéines et pour mettre en place la stratégie de recherche des glycoprotéines. Les protéines sont analysées après un temps de culture de deux semaines, ce qui correspond au début de la phase exponentielle de

croissance. Les surnageants de culture ont été obtenus après double filtration des cultures sur filtre de 0.22 μm , par Chia-Fang Liu, post-doctorante dans l'équipe de G. Puzo.

II. Conditions analytiques

Dans le premier milieu de culture 7H9 ADC, l'albumine est présente à une concentration de 2.5 g/L, ce qui compromet l'analyse protéomique réalisée par la suite pour la recherche de glycoprotéines. En effet, au cours d'une analyse protéomique, les différents peptides tryptiques de cette protéine présente en large excès vont être souvent séquencés au sein du spectromètre de masse, au détriment des peptides des autres protéines. Un nombre moindre de protéines identifiées va alors être obtenu.

Nous avons alors choisi d'utiliser une stratégie d'égalisation des protéines (réduction de la gamme de concentration des protéines), la stratégie Proteominer ((Guerrier, Thulasiraman et al. 2006) figure 30). Les billes Proteominer sont des billes sur lesquelles sont greffées des millions d'hexapeptides obtenus par chimie combinatoire. Chaque protéine de l'échantillon de départ va trouver un nombre limité d'hexapeptides pour lesquels elle va avoir une affinité. Ainsi, une protéine en large excès va rapidement saturer ses sites de fixation et, par lavage, la fraction non retenue de cette protéine va être éliminée.

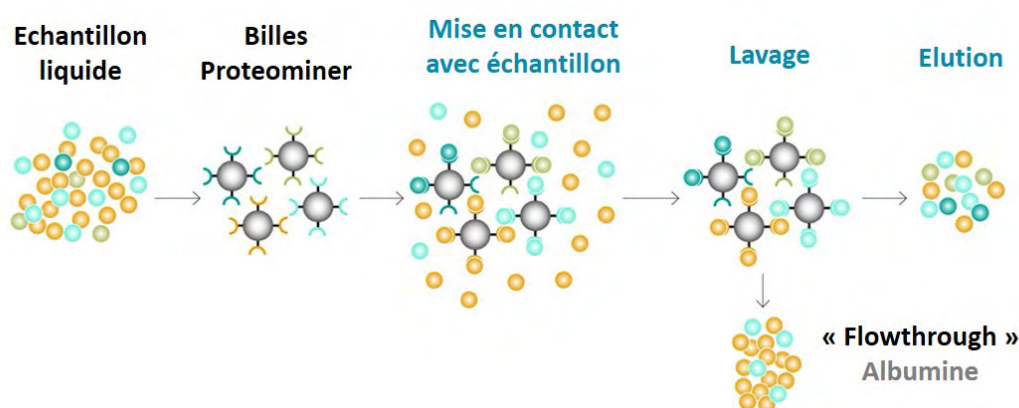


Figure 30: La stratégie Proteominer pour l'égalisation des protéines. L'échantillon d'intérêt est mis en présence des billes sur lesquelles sont greffées des millions d'hexapeptides. Après une étape de lavage pour éliminer les protéines qui n'ont pas trouvé de sites de fixation, les protéines fixées aux billes sont ensuite éluées grâce à un tampon de Laemli. Ce système permet alors de réduire la concentration relative d'albumine dans le milieu de culture.

Afin de faire subir aux protéines provenant des deux milieux de culture le même traitement biochimique, une étape d'égalisation a également été réalisée sur les protéines provenant du milieu de culture 7H9 DC.

Enfin, les protéines provenant des deux milieux de culture 7H9 DC et 7H9 ADC, sans passer par une étape d'égalisation, ont également été analysées. En effet, nous voulions analyser,

pour les deux milieux, une condition contrôle sans égalisation à comparer à la même condition avec égalisation. De plus, nous voulions également évaluer l'apport d'une stratégie d'égalisation des protéines sur l'enrichissement en glycoprotéines, souvent minoritaires. En comparant la liste des glycoprotéines obtenues après analyse des protéines provenant du milieu de culture 7H9 ADC avec ou sans étape d'égalisation, il nous était possible d'évaluer ce point.

Pour les échantillons correspondant aux milieux 7H9 DC et 7H9 ADC avec stratégie d'égalisation, les protéines présentes dans le surnageant de culture ont été mises en présence des billes Proteominer. Les protéines éluées des billes ont été dosées et 100 µg de protéines ont été fractionnées sur un gel 1D SDS-Page (figure 31). Pour les échantillons correspondant aux milieux 7H9 DC et 7H9 ADC sans stratégie d'égalisation, les protéines ont directement été dosées et fractionnées sur gel, de la même façon.

Sur chaque piste du gel, 8 bandes ont été découpées. Les protéines ont ensuite été digérées en gel par la trypsine. Chacune des fractions peptidiques ainsi obtenues a ensuite été analysée en nanoLC-MS/MS sur l'Orbitrap Velos (Thermo Fisher Scientific), avec un temps de séparation de 160 minutes. Ainsi, l'analyse complète d'une piste a été réalisée en 21 heures environ.

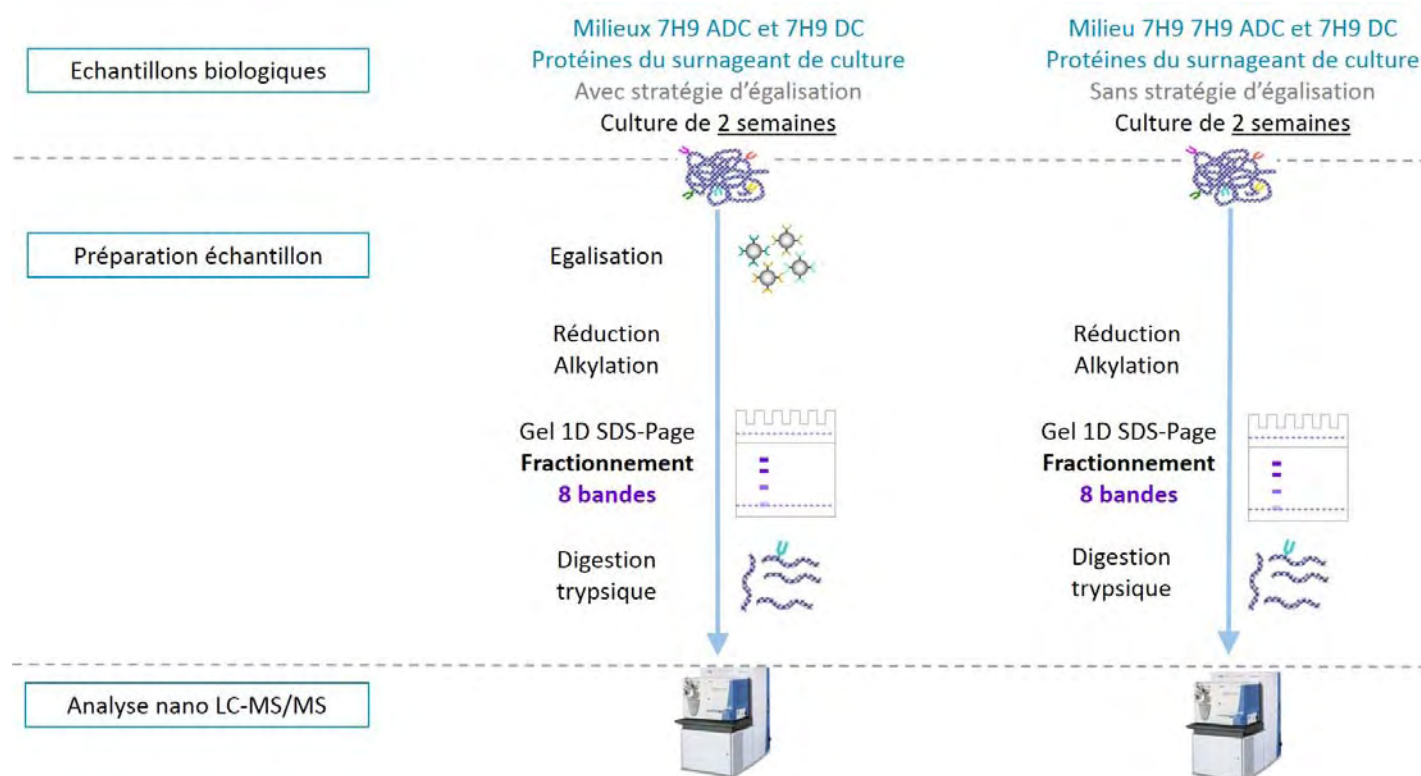


Figure 31: Stratégie d'analyse des protéines du surnageant de culture des milieux 7H9 ADC et 7H9 DC avec ou sans stratégie d'égalisation. Après une éventuelle égalisation sur les billes Proteominer, les protéines sont réduites et alkylées, fractionnées sur un gel SDS-Page. 8 bandes sont découpées sur chaque piste, et la digestion in gel des protéines à la trypsine est réalisée. Les peptides sont ensuite analysés par nanoLC-MS/MS.

III. Evaluation de l'apport d'une méthode d'égalisation sur l'identification des protéines

L'apport de l'égalisation dans le cas du milieu 7H9 ADC est bien visible sur le gel SDS-Page : sur la piste n°1 du gel (figure 32), la bande correspondant à l'albumine est très intense. Sur la piste n°2 après égalisation, cette bande est moins intense et la répartition des autres protéines le long de la piste est plus homogène.

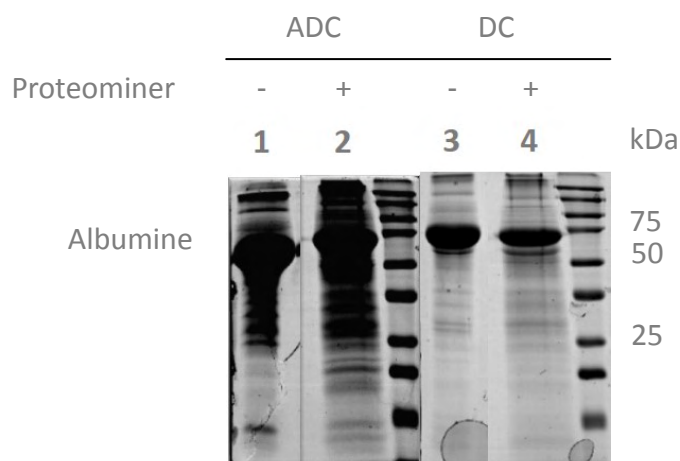


Figure 32: Gels SDS-Page des protéines des milieux 7H9 DC et 7H9 ADC, avec ou sans étape d'égalisation. Piste 1 : milieu 7H9 ADC sans étape d'égalisation – Piste 2 : milieu 7H9 ADC avec étape d'égalisation – Piste 3 : milieu 7H9 DC sans étape d'égalisation – Piste 4 : milieu 7H9 DC avec étape d'égalisation

En revanche, l'apport de l'égalisation dans le cas du milieu 7H9 DC est moindre. En effet, la répartition des protéines sur la piste 4 (milieu 7H9 DC avec étape d'égalisation) ne semble que légèrement plus homogène par rapport à la piste 3 (milieu 7H9 DC sans étape d'égalisation). Cette stratégie d'égalisation semble peu intéressante lorsqu'une protéine n'est pas présente en large excès.

A la suite de l'interrogation des banques de données par Mascot à partir des données brutes, afin d'obtenir le nombre de protéines totales validées, une étape de validation des protéines a été faite, avec un FDR de 5% (voir introduction première partie, paragraphe C.I). Les listes des protéines ainsi validées, pour chacune des expériences, ont été comparées.

Pour les deux échantillons correspondant aux milieux de culture 7H9 ADC et 7H9 DC sans stratégie d'égalisation, plus de 300 protéines ont été identifiées et validées (figure 33). Dans une étude de 2012, 422 protéines ont été identifiées dans le surnageant de culture du bacille tuberculeux par analyse LC-MS/MS (Bell, Smith et al. 2012, Smith, Sweredoski et al. 2014). L'étape d'égalisation des protéines du milieu 7H9 DC permet alors d'atteindre ce

nombre, avec un gain de plus de 100 protéines par rapport à l'analyse du même milieu sans étape d'égalisation.

Dans le cas du milieu 7H9 ADC, sans étape d'égalisation, 349 protéines sont validées, contre 904 avec égalisation. Le gain en nombre de protéines validées est donc considérable pour ce milieu 7H9 ADC pour lequel l'albumine est présente en large excès.



Figure 33: Nombre de protéines validées selon la condition étudiée. 7H9 ADC PM = milieu 7H9 ADC avec égalisation, 7H9 DC PM = milieu 7H9 DC avec égalisation

Il a été montré dans la littérature que cette technique d'égalisation permet l'identification d'un plus grand nombre de glycopeptides dans le cas de protéines N-glycosylée (Bandhakavi, Van Riper et al. 2011). Afin d'évaluer l'apport de cette stratégie pour un enrichissement en O-glycoprotéines, nous avons alors poursuivi l'étude par une recherche des glycoprotéines du bacille.

IV. Le retraitement bio-informatique des données pour l'identification de glycoprotéines

Sur les données acquises, un outil bioinformatique développé au laboratoire, « Neutral Loss », a été utilisé afin de « filtrer » les données et rechercher les glycopeptides contenus dans chacun des échantillons.

En effet, les moteurs de recherches actuels tels que Mascot ou OMSSA ne sont pas spécifiquement dédiés à la recherche de protéines modifiées.

1. La recherche des protéines glycosylées avec le moteur de recherche Mascot

Au cours d'une interrogation des bases de données avec Mascot, il est possible de rechercher des peptides possédant certaines PTMs que l'utilisateur sélectionne. Il est alors possible de rechercher des peptides glycosylés. En revanche, il est impossible de maîtriser la qualité des spectres glycosylés sélectionnés par Mascot.

Ainsi, Il n'est pas possible de choisir la tolérance de masse pour la recherche de l'ion provenant d'une éventuelle perte de neutre caractéristique des spectres de peptides glycosylés (figure 34). De plus, il n'est pas non plus possible de déterminer une intensité minimale pour celui-ci. Des spectres vont alors être interprétés comme spectres de peptides glycosylés alors que l'ion provenant de la perte de neutre considéré à une intensité relativement faible, et peut même faire partie du bruit de fond.

De plus, un ion de fragmentation y ou b peut être interprété comme un ion provenant de la perte de neutre. Ceci peut alors entraîner une fausse attribution de la séquence peptidique.

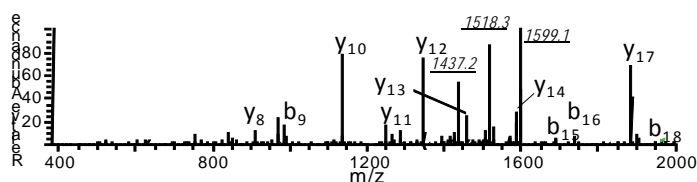


Figure 34: Spectre de masse d'un peptide tri-glycosylé vrai-positif. Au cours de la fragmentation du peptide avec un mode de fragmentation CID, il y a formation au sein du spectre de 3 ions intenses provenant des 3 pertes d'hexoses successives à partir de l'ion parent.

Le logiciel « Neutral Loss » a alors été créé pour une meilleure sélection des spectres de peptides glycosylés putatifs. Ce logiciel est complémentaire à l'utilisation de Mascot. En effet, « Neutral Loss » va permettre de filtrer la liste des spectres attribués à des peptides glycosylés par Mascot.

2. Principe du logiciel « Neutral Loss »

Le logiciel « Neutral Loss » a initialement été développé, avant ce travail de thèse, en travaillant avec des données issues de l'analyse LC-MS/MS des protéines du surnageant de *Mycobacterium smegmatis*, souche à croissance rapide et non pathogène.

Le principe de ce logiciel est basé sur la caractéristique que possèdent les spectres MS/MS de peptides glycosylés (voir introduction première partie, paragraphe C.I.) : présence d'un ou de plusieurs ion(s) intense(s) au sein de chacun des spectres correspondant à la masse de l'ion parent soustraite de la masse d'un ou de plusieurs hexoses. Le logiciel recherche, parmi

tous les spectres MS/MS acquis au cours d'une analyse, les spectres possédant cette caractéristique.

A partir de 100µg de protéines du surnageant de culture de *Mycobacterium smegmatis* déposées sur un gel SDS-Page et analysées par nanoLC-MS/MS, dix glycoprotéines avaient été identifiées (tableau 2).

Le développement de ce logiciel a ensuite été poursuivi au cours de ce travail de thèse avec l'analyse des protéines de *Mycobacterium tuberculosis*.

Protein description	Access no.*	Gene no.	MM [†]	Peptide sequence	Hexose no.
Fascin domain protein	A0R2Q4	5196	3020.3	(A)EPETTTEAEPTEIPDPQGPQC [‡] DDFK(K)	1-3 Hex
			3148.4	(A)EPETTTEAEPTEIPDPQGPQC [‡] DDFKK(A)	1-3 Hex
ABC-type amino acid transport system secreted component	A0QXB0	3235	1867.8	(S)GEGGGESSPESTQAASGAK(V)	1-2 Hex
			1954.8	(A)SGEGGGESSPESTQAASGAK(V)	1-2 Hex
			2025.9	(C)ASGEGGGESSPESTQAASGAK(V)	1-3Hex
Putative protein	A0R6M4	6601	2574.3	(K)STTPPTPLPPLPAEVGGSAAK(A)	3-5 Hex
Putative uncharacterized protein	A0QYI9	3674	2777.3	(K)VILMLDC [§] TEAAQQAQDTAVSGGPR(V)	1 Hex
			3068.5	(R)VGSN [§] GVTTVTPTPVPIAGAPGAGTPPPA(-)	3-4 Hex
Putative uncharacterized protein	A0QRP2	1188	1876.9	(R)AQPVTPTQSDTVPGR(Y)	2 Hex
			3702.9	(R)YLPVPVPEHPDVALPVTVLTTPPAQPVAPNTR(T)	2 Hex
Immunogenic protein MPT63	A0R3B5	5412	1904.0	(T)TTVPAPAPTTAPVSHGPAA(-)	1 Hex
			2389.2	(K)DLLVWDKPAASATAPSGTGQSR(P)	1 Hex
			4227.9	(R)ATDLPAATAEAETEATATDAETTPVTEETPAAPAEPTTPV(S)	1 Hex
Putative protein	A0QSU1	1600	1994.0	(K)LQTLTTTVATVNDEAQK(Q)	1 Hex
Immunogenic protein MPT63	A0QQP4	828	1492.7	(R)PAATGSGTSASTPAR(T)	1 Hex
Putative protein	A0QVP4	2645	3127.5	(K)WPVIETTDPKPFDPK [‡] NDIPIDVIER(I)	1 Hex
Alanine and proline-rich secreted protein apa	A0QYD3	3618	2119.0	(R)PGVGVPVPTDAPPEMMPPA(-)	1 Hex

Boldface T and S correspond to potential glycosylation sites according NetOGlyc (v3.1) predictions.

*UniProtKB accession number.

[†]Experimental mass of the glycosylated peptide of lower glycosylation degree.

[‡]Carboxymethylated cysteine.

[§]Cysteine acrylamide adduct.

[¶]Deamidation.

Tableau 2: Liste des glycoprotéines de *Mycobacterium smegmatis* identifiées grâce au logiciel « Neutral Loss ».

3. Paramétrisation du logiciel

Les données d'entrée de l'outil « Neutral Loss » sont contenues dans les « peak list » enregistrées au cours des analyses nanoLC-MS/MS. Ces « peak list » renferment la liste de toutes les masses des ions contenus dans les spectres MS et MS/MS enregistrés, avec les intensités correspondantes. Plusieurs « peak list » peuvent être entrées de façon simultanée dans l'outil « Neutral Loss » (figure 35).

Certains paramètres de recherche sont à définir : la précision de masse sur la perte de masse d'un ou de plusieurs hexoses depuis l'ion parent, l'intensité relative minimale que peuvent posséder les ions provenant de la perte de neutre au sein du spectre et le rang maximum de chacun de ces ions parmi tous les autres ions du spectre. Dans le cas d'un peptide avec un seul mannose, l'ion unique provenant de la perte de neutre à partir de l'ion parent sera un des ions les plus intenses du spectre. Comme valeur d'intensité relative minimum à entrer

en tant que paramètre avant de lancer l'outil « Neutral Loss », nous pourrions choisir une valeur arbitraire relativement élevée, de 70% par exemple.

En revanche, dans le cas d'un peptide avec quatre mannoses, la dernière perte de neutre à partir de l'ion parent aura une intensité relativement faible et pourra être parmi les ions les moins intenses du spectre. Nous avons donc choisi une valeur d'intensité minimale de 20% et un rang maximal de 15, pour ne pas risquer d'éliminer d'éventuels peptides multiglycosylés vrais positifs au cours du crible. En ce qui concerne la précision sur la masse de la perte de neutre, nous l'avons choisi de 500 ppm.

En sortie de cette recherche, un fichier contenant la liste des spectres MS/MS possédant une signature de peptide glycosylé est renvoyée, avec certaines informations dont l'intensité au sein des spectres de chacun des ions provenant d'une éventuelle perte de neutre à partir de l'ion parent et le rang de ceux-ci au sein du spectre.

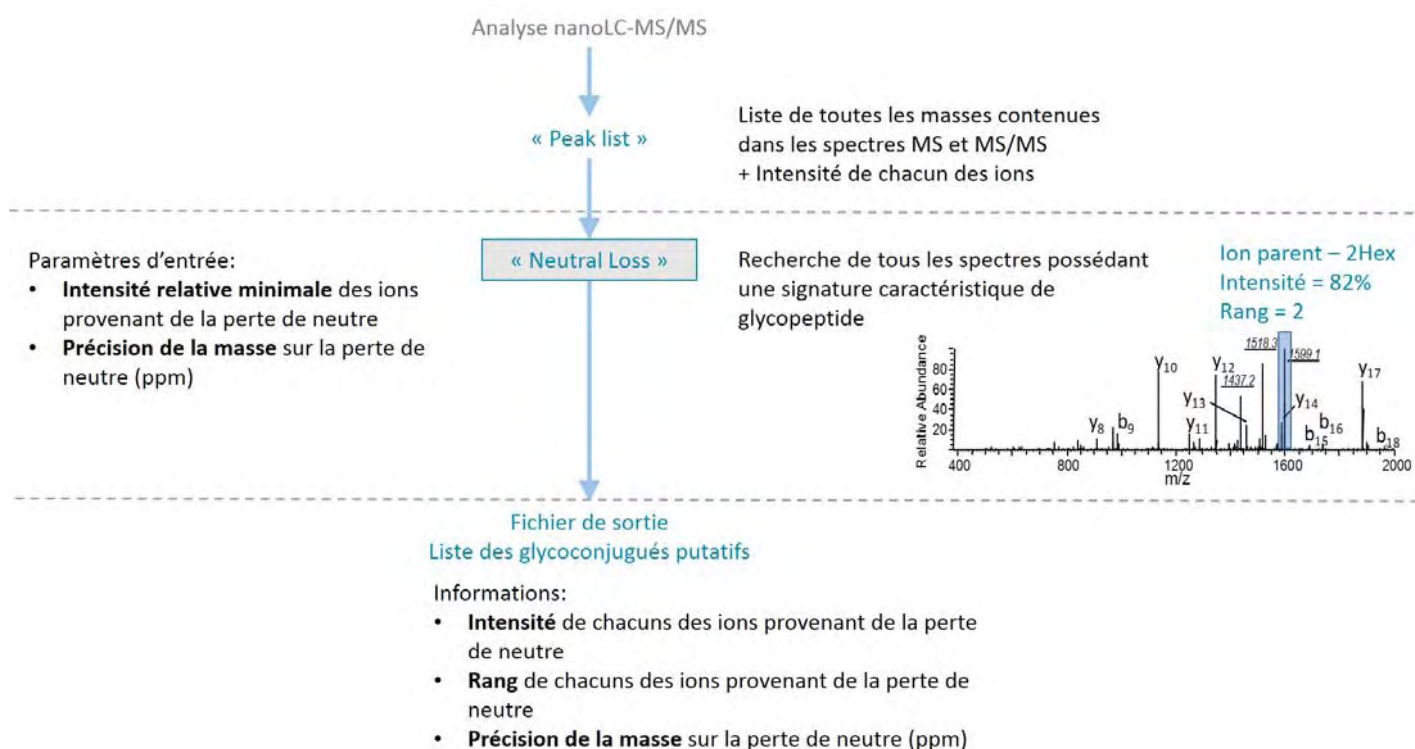


Figure 35: Workflow d'utilisation du logiciel Neutral Loss

4. Retraitement des données issues du logiciel « Neutral Loss »

La liste de spectres MS/MS que « Neutral Loss » considère comme spectres possédant la signature caractéristique de peptides glycosylés est une liste de glycoconjugués putatifs. En effet, dans les surnageants de culture analysés, beaucoup de glycolipides ou autres

molécules glycosylées sont présents et peuvent être fragmentés dans le spectromètre de masse de la même façon que les glycopeptides.

Le logiciel Mascot permet alors ensuite de confronter les spectres MS/MS expérimentaux à des spectres MS/MS théoriques de peptides de protéines renfermées dans des bases de données, et donc d'obtenir une liste de glycopeptides putatifs.

A la suite de l'interrogation des données par Mascot, de façon générale au cours d'une analyse LC-MS/MS, les spectres MS/MS sont classés en trois catégories (figure 36):

- **Catégorie 1** : certains spectres MS/MS ne correspondant à aucune interprétation peptidique. Ces spectres peuvent alors correspondre à des composés autres que des peptides ou être de très mauvaise qualité et peu informatifs.
- **Catégorie 2** : une séquence peptidique est attribuée à certains spectres MS/MS mais les peptides correspondant ne sont pas attribués à une protéine. Dans ce cas, un seul peptide appartient à la protéine correspondante et celle-ci n'est alors pas validée par Mascot.
- **Catégorie 3** : les spectres MS/MS pour lesquels une interprétation de séquence peptidique est donnée, ainsi que l'appartenance à une protéine.

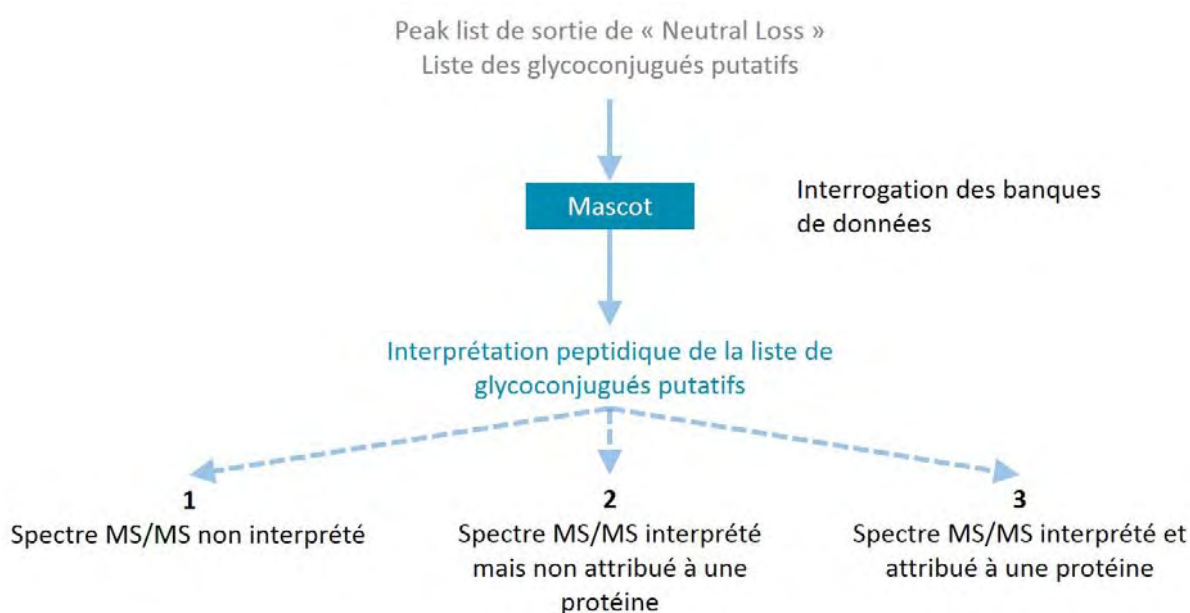


Figure 36: A la suite de la recherche des interprétations peptidiques des spectres MS/MS enregistrés au cours d'une analyse nanoLC-MS/MS par Mascot, les spectres peuvent être classés en 3 catégories : les spectres pour lesquels aucune interprétation peptidique n'a été donnée, les spectres pour lesquels une interprétation peptidique a été donnée mais pour lesquels aucune protéine n'a été attribuée, les spectres pour lesquels une interprétation peptidique ainsi que l'appartenance à une protéine ont été donnés.

Afin d'établir la liste des glycoprotéines du bacille, nous nous sommes alors intéressés aux spectres MS/MS retenus par le logiciel « Neutral Loss » et classés dans la catégorie n°3 après interrogation des bases de données avec Mascot.

Pour valider un candidat glycopeptide, il faut qu'un (ou plusieurs) spectre MS/MS de ce peptide soit d'une part sélectionné par l'outil « Neutral Loss » et d'autre part qu'une séquence peptidique de peptide glycosylé lui ait été attribuée par Mascot. L'outil « Neutral Loss » permet alors bien de filtrer les interprétations peptidiques de Mascot.

De plus, plusieurs attributions de séquences peptidiques peuvent être attribuées à un même spectre MS/MS, nous n'avons alors considéré que les attributions de rang 1.

V. Résultats

1. Apport de l'outil bioinformatique Neutral Loss pour la sélection des spectres de glycopeptides putatifs

Après analyse LC-MS/MS de chacune des huit fractions peptidiques correspondant à une piste du gel 1D SDS-Page, 100 000 spectres MS/MS environ ont été enregistrés pour chacun des 4 échantillons biologiques (7H9 ADC, 7H9 DC, 7H9 ADC PM avec égalisation Proteominer et 7H9 DC PM avec égalisation Proteominer).

Sur ces 100 000 spectres, l'outil Neutral Loss a sélectionné environ 1500 spectres possédant une signature de peptide glycosylé, ce qui représente seulement 1.5% des spectres (tableau 3).

Après confrontation de ces données brutes aux données obtenues avec Mascot, un faible nombre de spectres MS/MS appartiennent à la catégorie n°3. Parmi ces spectres, peu ont, selon Mascot, une attribution de séquence de peptide glycosylé (1% de nombre total de spectres environ).

Enfin, il est nécessaire que le nombre d'unités sucre par peptide détectés par Mascot sur un peptide soit identique à celui détecté par « Neutral Loss » dans le spectre MS/MS correspondant. En effet, selon le nombre de pertes de neutres que Mascot va détecter au sein d'un spectre MS/MS, une masse correspondant à la masse de ces unités osidiques va être soustraite de la masse expérimentale mesurée du précurseur. C'est à cette masse recalculée qu'une séquence peptidique nue va être recherchée dans les bases de données et attribuée (et le nombre d'unités sucre va être indiqué). Il est alors impératif que le degré de glycosylation détecté soit correct, et confronter ce nombre détecté par « Neutral Loss » avec celui détecté par Mascot permet de limiter les erreurs.

En ne considérant que les spectres MS/MS correspondant alors à ce critère, seulement quelques dizaines de spectres MS/MS peuvent correspondre à des glycopeptides putatifs

(tableau 3). Parmi ces spectres, certains spectres correspondent aux mêmes peptides, et seulement moins de 10 glycoprotéines sont ainsi identifiées, selon les conditions.

	7H9 ADC	7H9 DC	7H9 ADC PM	7H9 DC PM
Nombre protéines totales validées	349	337	904	460
Nombre de spectres MS/MS enregistrés	117148	121131	103711	104450
Spectres avec signature de glycoconjugués (Neutral Loss)	1456	1570	1293	1567
% par rapport au nombre total de MS/MS	1,24	1,30	1,25	1,50
Spectres de glycoconjugués putatifs de catégorie 3, avec interprétation de rang 1	70	394	96	174
Spectres de glycoconjugués putatifs avec attributions de séquences de peptides glycosylés	12	43	6	32
Spectres de glycoconjugués putatifs avec un même nombre d'hexoses détectés par Mascot et par neutral Loss	8	34	5	25
Spectres de glycoconjugués putatifs uniques avec un même nombre d'hexoses détectés par Mascot et par neutral Loss	3	4	7	8
Nombre de protéines glycosylées putatives uniques	3	3	6	7

Tableau 3: Utilisation de l'outil "Neutral Loss" pour la sélection des spectres de peptides glycosylés putatifs – milieu 7H9

2. Glycoprotéines identifiées

Douze protéines glycosylées putatives ont ainsi été identifiées, sur l'ensemble des quatre conditions (tableau 4).

Certaines des protéines sont identifiées avec un seul peptide glycosylé, et d'autres sont soit identifiées avec plusieurs peptides possédant la même séquences peptidique mais différent degrés de glycosylation, soit avec des peptides glycosylés de séquences différentes. Ainsi, deux glycoprotéines sont identifiées avec deux peptides glycosylés différents, la lipoprotéine Lpra et une protéine putative.

Dans le cas des peptides monoglycosylés, étant donné qu'il y a au sein du spectre un seul ion provenant de la perte du neutre à partir de l'ion parent, c'est pour ces peptides que le taux de faux positif risque d'être le plus important (ion au sein du spectre interprété comme un ion provenant d'une perte de neutre qui est en réalité un ion y ou b provenant de la fragmentation du peptide).

Afin d'essayer d'évaluer la fiabilité des résultats obtenus, nous nous sommes intéressés aux probabilités de glycosylation des peptides d'intérêt fournies par le logiciel NetOGlyc. Ce

programme est un logiciel de prédiction des sites de glycosylation de *O*-glycoprotéines, basé sur la nature des acides aminés qui composent les protéines et sur l'accessibilité des sites potentiels de glycosylation. Ces règles ont été établies suite à l'étude de protéines *O*-glycosylées connues et caractérisées (Julenius, Molgaard et al. 2005, Steentoft, Vakhrushev et al. 2013).

Sur les douze protéines identifiées, seulement six possèdent des glycopeptides ayant de bonnes probabilités de glycosylation selon NetOGlyc. De plus, aucune des glycoprotéines putatives identifiées n'a été décrite comme glycosylée dans la littérature.

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

Gene Rv	Protéine	Expérience		AA	PTM	Degré glyco	Queries
Rv0315	Beta-1,3-glucanase putative	DC	APAGPT(0,92)PAPAAPAAATGGLLFHDEFDGPAGSVDPDSK	A		1	4
Rv0468	3-hydroxybutyryl-CoA dehydrogenase FADB2 putative	DC PM	VAAATKQQRVL	K		1	1
Rv1079	Cystathionine gamma-synthase MetB (Cgs)	ADC PM	FGGMVSRLHGGR	G	O (M)	1	1
Rv1270c	Lipoprotein putative LPRA	DC	AAT(0,69)ASNGDAAMLLK	T	D (N)	1	1
		ADC PM	AS(0,81)DT(0,77)AAT(0,69)ASNGDAAMLLK	K	D (N)	1	1
		DC PM					1
		DC					4
		ADC					1
		DC		K	D (N)	2	2
		DC PM		K	D (N); O (M)	1	1
		DC					2
		ADC					1
Rv2145c	Antigène sécrété Wag31	ADC	ADGMLADAQARS(0,60)ESQ	R		1	1
Rv2385	Acyl-[acyl-carrier protein] desaturase DESA2 putative (Acyl-[ACP] desaturase) (Stearoyl-ACP desaturase)	DC PM	ANLVTHCLDYTR	F		1	1
Rv2799	Protéine membranaire putative	ADC PM	SPIVAT(0,73)T(0,84)DPS(0,53)PFDPCR	K	C (C)	1	1
		DC PM				2	17
		DC				2	12
		ADC				1	1
Rv2831	Enoyl-CoA hydratase à courte chaîne	DC PM	CAGLDLKELGDTQLPDISP	F		1	1
Rv2845c	Ligase proline--tRNA ligase	ADC PM	GQTRELSIDADLAGQVLAARAEAG	S		1	1
Rv2971	Oxidoreductase putative	ADC PM	LLDNPTVTS(0,64)IASEYVKTPAQVLLRWN	R		1	1
Rv3255c	Isomerase mannose-6-phosphate	DC PM	ANSDNVLRGGLTPK	M		1	1
Rv3491	Protéine putative non caractérisée	ADC PM	SLQLIGPPSPVQR	F		1	2
		DC PM		F		1	9
		DC PM	QPFSQLIGPPSPVQR	K		1	2
		ADC PM		K		1	1

Tableau 4: Liste des glycoprotéines de *Mycobacterium tuberculosis* putatives identifiées après analyse des protéines des surnageants des milieux 7H9 ADC, 7H9 DC, 7H9 ADC PM et 7H9 DC PM. AA = acide aminé en amont du peptide, O(M) = oxydation de la méthionine, D(N) = déamidation de l'asparagine, C(C) = carbamidométhylation de la cystéine. En bleu, probabilité de glycosylation des sérines et thréonines selon le logiciel NetOGlyc.

VI. Conclusion

Douze protéines glycosylées ont pu être identifiées au cours de l'analyse des protéines du surnageant de culture en milieu 7H9 (conditions 7H9 ADC, 7H9 ADC PM, 7H9 DC, 7H9 DC PM confondues). Les glycoprotéines de *Mycobacterium tuberculosis* décrites dans la littérature (Apa, SodC, MPB83, LpqH et LprG) n'ont d'ailleurs pas été identifiées dans cette première étude, et peu d'entre elles possèdent des peptides glycosylés putatifs qui ont de bonnes probabilités de glycosylation selon NetOGlyc.

De plus, le procédé d'égalisation des protéines sur billes Proteominer ne semble pas permettre un enrichissement en glycoprotéines. En effet, le nombre final de spectres retenus par « Neutral Loss » comme spectres de peptides glycosylés putatifs n'est pas plus important dans le cas des conditions avec égalisation par rapport aux mêmes conditions sans égalisation (tableau 3).

Afin d'améliorer ce résultat, nous avons alors ensuite décidé d'évaluer l'apport de l'utilisation de méthodes d'enrichissement pour cette recherche de glycoprotéines.

C. Essais d'enrichissement

Nous avons testé deux approches : l'enrichissement des glycoprotéines ou glycopeptides sur lectine Concanavalline A et la chromatographie hydrophile HILIC.

I. Enrichissement par affinité à la Concanavalline A

Nous avons évalué l'apport de l'affinité à la concanavalline A pour un enrichissement en glycoprotéines ou glycopeptides d'un échantillon complexe.

1. Enrichissement au niveau protéique

Nous avons alors tout d'abord utilisé l'enrichissement au niveau protéique, à partir de 1mg de protéines totales du surnageant d'une culture de la souche sauvage en milieu sauton de 6 semaines. Après une incubation de 2h des protéines avec la phase de lectine, la phase est récupérée, lavée et les protéines ayant une affinité avec celle-ci sont éluées grâce à un compétiteur.

Nous avons ensuite déposé 90% de cette fraction d'élution sur un gel SDS-Page avec fractionnement afin de visualiser d'éventuelles bandes correspondant à des protéines

putatives glycosylées (figure 37). Une bande apparaît, avec un poids moléculaire apparent de 37kDa environ, sur la piste des protéines ayant une affinité pour la lectine.

Les 10% restant de cette fraction ont été déposés sur un gel SDS-Page sans fractionnement pour analyse nanoLC-MS/MS. La digestion des protéines a été réalisée *in gel* et les peptides ont ensuite été analysés. Cette stratégie permet d'obtenir des informations sur l'ensemble des protéines présentes dans l'échantillon en une seule analyse. Cette stratégie est utile lorsque peu de protéines sont présentes dans l'échantillon, ce qui est ici le cas, d'après le gel SDS-Page réalisé.

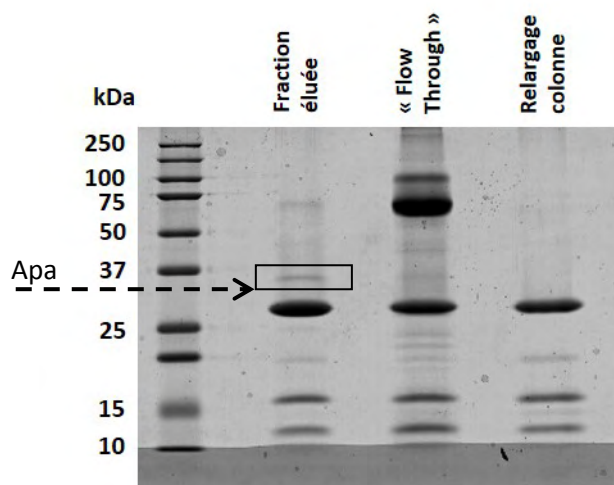


Figure 37: Gel SDS-Page réalisé avec la fraction protéique enrichie sur concanavaleine A (première piste), la fraction non enrichie (deuxième piste) et du tampon de lavage passé sur la colonne afin d'identifier les protéines relarguées par la colonne.

Plus de 10000 spectres MS/MS sont enregistrés au cours de cette analyse. Plus de 12% de ces spectres sont sélectionnés par Neutral Loss comme spectres de glycoconjugués putatifs. Cependant, à un grand nombre de ces spectres, aucune attribution de séquence peptidique n'est donnée par Mascot. Seulement 2 glycoprotéines sont validées : Apa (dont la bande est visible sur le gel) et une protéine putative (tableau 5). En effet, la concanavaleine A ne permet pas d'enrichir de façon spécifique en glycoprotéines ou glycopeptides, tous les composés mannosylés ayant une affinité pour cette lectine (Becker, Reeke et al. 1975). Beaucoup de spectres MS/MS enregistrés dans cette étude sont des spectres de composés multiglycosylés autre que des peptides (glycopeptides par exemple) car Mascot ne leur attribue pas de séquences peptidiques, et peuvent alors correspondre à d'autres glycoconjugués. De plus, l'affinité de la lectine pour les glycoprotéine ou glycopeptides demeure faible.

Gene Rv	Protéines	Peptides glycosylés	Degré glyco	Queries
3705a	Protéine putative riche en proline	GPTGPGPTTAPARP	3	1
		VGPTGPGPTTAPARP	2	1
		AVGPTGPGPTTAPARP	2	2
		GGPAVGPTGPGPTTAPARP	2	1
1860	Apa	APTPTTPTPQRTLPA	2	1
		GEVAPTPTTPTPQRTLPA	2	2

Tableau 5: Liste des protéines identifiées après enrichissement avec la concanavaleine A au niveau protéique

2. Enrichissement au niveau peptidique

Pour cet essai d'enrichissement sur lectine au niveau peptidique, deux échantillons de 20 µg de protéines totales du surnageant ont été déposées sur gel SDS-Page sans fractionnement sur deux pistes différentes et les protéines ont été dirigées par la trypsine en gel.

Une des deux fractions peptidiques a été incubée avec une phase de lectine selon le même protocole que pour l'enrichissement niveau protéique et les peptides ayant une affinité pour cette phase ont été élués et analysés en nanoLC-MS/MS. L'autre fraction est analysée sans enrichissement pour comparaison.

Plus de 8000 spectres ont été enregistrés au cours de la première expérience (enrichissement sur ConA), contre plus de 38000 pour la seconde expérience. Pour la première expérience avec enrichissement, 3.8% des spectres sont sélectionnés par « Neutral Loss », contre 3.13% pour l'expérience sans enrichissement, la différence dans le nombre de spectres de glycopeptides putatifs analysés ne semble alors pas significative à ce niveau. En revanche, après application des différents filtres, 0.2% des spectres MS/MS sont validés comme spectres de glycopeptides putatifs pour l'expérience avec enrichissement contre 0.02% pour l'expérience sans enrichissement.

Pour la première expérience, 2 glycoprotéines sont validées, la protéine membranaire codée par le gène Rv2799 et la beta-lactamase (tableau 6). Pour la deuxième expérience, sans enrichissement, aucune glycoprotéine n'a été identifiée.

Gene Rv	Protéines	Petides glycosylés	Degré glyco	Queries
2068c	Beta-lactamase	ARPASTTLPAGADLADR	2	2
		STTLPAGADLADR	2	2
		PASTTLPAGADLADR	2	1
2799	Protéine membranaire	SPIVATTDPSFDPCR	1	1
		SPIVATTDPSFDPCR	2	4

Tableau 6: Liste des protéines identifiées après enrichissement avec la concanavaleine A au niveau peptidique

Compte-tenu du fait que cette expérience nous a permis d'identifier 2 glycoprotéines à partir de 20 µg de protéines digérées seulement, nous avons essayé de poursuivre ces essais et de répéter l'expérience avec des réplicats biologiques. En effet, nous voulions essayer d'identifier d'autres protéines et également tester la reproductibilité de ce protocole afin de l'utiliser sur des digestats peptidiques obtenus après digestion de plusieurs bandes d'un gel 1D avec fractionnement des protéines. Cependant, après plusieurs essais, nous ne sommes pas parvenus à répéter ces résultats et nous n'avons pu identifier aucune glycoprotéine. Il semblerait alors que ce type d'expérience ne soit pas reproductible, nous n'avons donc pas envisagé d'utiliser cette méthode d'enrichissement.

II. Enrichissement par chromatographie d'interaction hydrophile

De la même façon que précédemment, 20 µg de protéines totales du surnageant ont été digérées *in gel*. Séchés, le digestat a ensuite été repris dans une solution acide avec un fort taux d'acétonitrile, puis déposé sur une phase hydrophile de type ZIC-HILIC. Les peptides hydrophiles sont ensuite élués avec une solution acide riche en eau.

Après analyse nanoLC-MS/MS de cette fraction éluée, environ 1.2% des spectres sont sélectionnés par Neutral Loss. En revanche, une très faible proportion de ces spectres possède une attribution de séquence peptidique après interrogation Mascot et, après application des différents filtres, aucune glycoprotéine n'est validée.

Cette méthode d'enrichissement sur phase chromatographique hydrophile n'est pas spécifique des glycopeptides. En outre, dans le cadre de notre étude, le caractère hydrophile apporté par l'oligosaccharide porté par les peptides peut ne pas prévaloir sur le caractère hydrophile ou hydrophobe apporté par les acides aminés du peptide. En effet, dans le cas d'un peptide di- ou tri-mannosylé, l'hydrophilie apportée par les mannoses est faible, contrairement au cas des N-glycopeptides qui portent le plus souvent des motifs glycosidiques d'une masse importante (acides sialiques par exemple).

Du fait des résultats obtenus avec ces deux méthodes d'enrichissement, nous avons décidé de ne pas les utiliser dans le cadre de notre étude et de poursuivre celle-ci avec un développement plus approfondi de l'utilisation du programme « Neutral Loss ».

D. Recherche des glycoprotéines sécrétées de *Mycobacterium tuberculosis* – Milieu de culture Sauton

I. Echantillons biologiques et conditions analytiques

Comme il a précédemment été évoqué, pour s'affranchir des problèmes liés à l'albumine présente en excès dans le milieu de culture 7H9, des essais de culture des trois souches du bacille (sauvage, mutante $\Delta Rv1002c$ et complémentée $\Delta Rv1002c :rv1002c$) avaient été réalisés. Le milieu de culture Sauton est un milieu composé d'acides aminés, de sels de sodium, magnésium, fer et potassium, et de glycérol comme source de carbone. C'est à partir de l'analyse de surnageants de culture provenant de ce milieu de culture que Apa, SodC et MPB83, totalement caractérisées dans la littérature ont été analysées. De plus, dans ces études, le temps de culture de la bactérie auquel ont été récupéré les surnageants était un temps de culture relativement long de plus de six semaines, correspondant à la phase exponentielle.

La croissance des trois souches en milieu liquide est similaire. Nous avons décidé de poursuivre la recherche des glycoprotéines de *Mycobacterium tuberculosis* avec ce nouveau milieu.

De la même façon que pour l'analyse des protéines du surnageant du milieu de culture 7H9, les protéines présentes dans le surnageant de culture du milieu Sauton ont été dosées, 100 μ g de protéines ont été fractionnées sur un gel 1D SDS-Page (figure 38). Une bande très intense apparaît et correspond à une ou plusieurs protéine(s) d'un poids moléculaire apparent entre 50 et 75 kda, poids moléculaire des protéines de type Hsp (« Heat Shock Protein ») souvent présentes de façon abondante dans les surnageants de culture. La répartition des protéines le long de la piste du gel est similaire entre les deux répliquats biologiques.

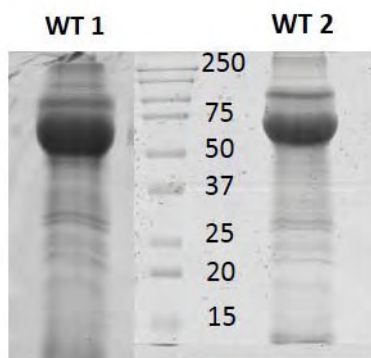


Figure 38: Gel SDS-Page réalisé avec les protéines du surnageant de culture des deux répliquats de la souche sauvage de *Mycobacterium tuberculosis*

De plus, contrairement à l'étude précédente, pour essayer d'obtenir un fractionnement plus important des protéines, nous avons décidé de découper 17 bandes par piste. En effet, découper un nombre plus important de bandes sur une piste d'un gel SDS-Page peut nous permettre d'obtenir une couverture analytique plus large et d'identifier un nombre global de protéines plus important, et donc un nombre de protéines glycosylées plus important également. Les protéines ont ensuite été digérées en gel par la trypsine et les fractions peptidiques ont été analysées en nanoLC-MS/MS sur l'Orbitrap Velos avec un temps de séparation de 160 minutes.

II. Retraitement des données

Comme il a déjà été évoqué précédemment, il est nécessaire que le nombre de pertes de neutre détectées par le logiciel « Neutral Loss » au sein d'un spectre soit identique à celui détecté par le moteur de recherche Mascot, pour une interprétation peptidique correcte.

Or, une tolérance de masse de 500 ppm sur ces pertes de neutre peut ne pas être suffisante pour ce but. Nous avons alors choisi d'élever cette tolérance à 800 ppm.

De plus, nous avons également choisi d'appliquer des filtres supplémentaires sur la liste de glycopeptides putatifs obtenue après interrogation Mascot de la liste de glycoconjugués putatifs de sortie de « Neutral Loss » (figure 39).

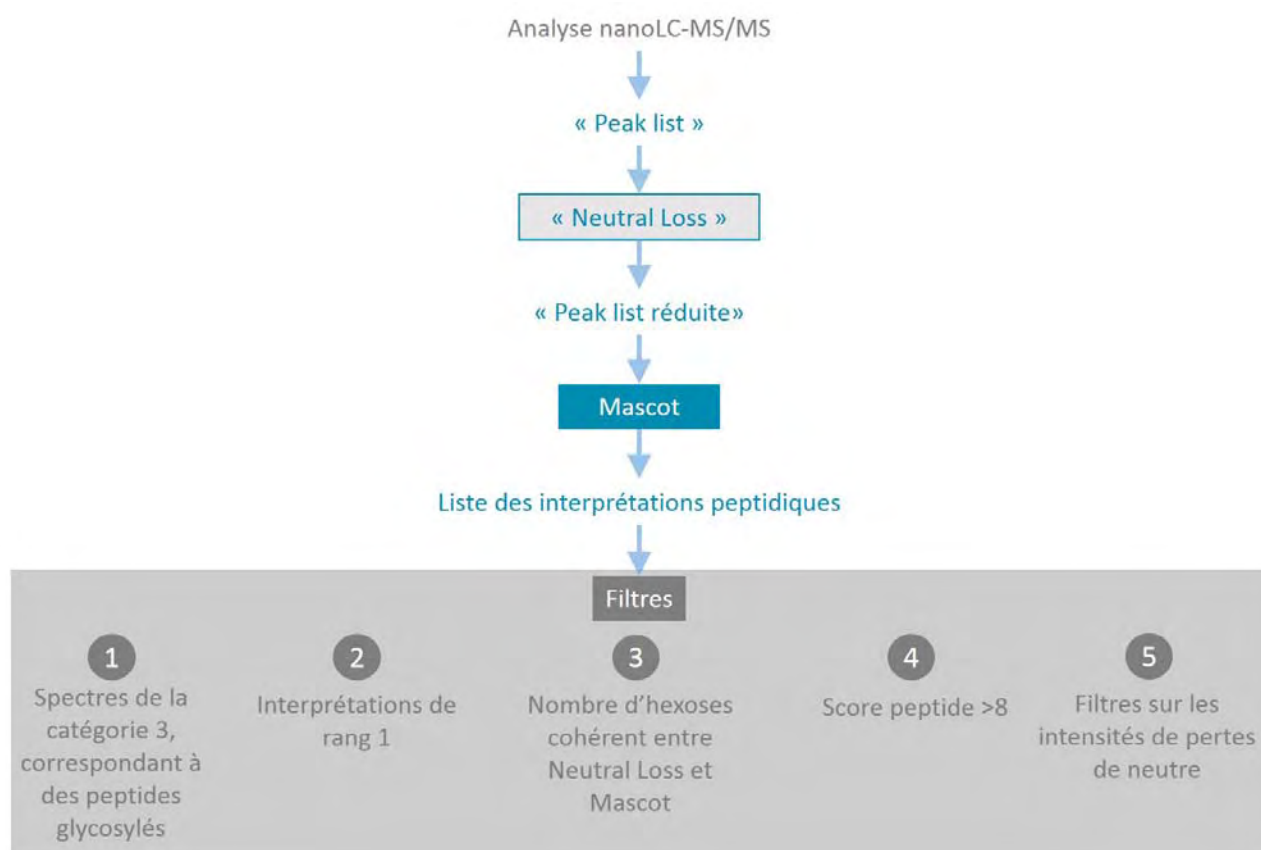


Figure 39: Nouveau workflow utilisé pour filtrer la liste des spectres MS/MS retenus par “Neutral Loss” comme spectres de glycopeptides putatifs

Premier filtre : Comme précédemment, après interrogation des banques de données par Mascot, seuls les spectres avec une interprétation de peptide glycosylé, et attribués à des protéines sont considérés.

Deuxième filtre : Seules les interprétations peptidiques de rang 1 sont retenues.

Troisième filtre : ne sont conservées que les interprétations peptidiques dont le nombre d'hexoses considéré par Neutral Loss est le même que celui considéré par Mascot.

Quatrième filtre : Seules les interprétations peptidiques de rang 1 avec un score supérieur à 8 sont conservées. En effet, ce score est une mesure du fait que cette interprétation soit due ou non au hasard. Nous avons arbitrairement choisi une valeur seuil de 8.

Cinquième filtre : un filtre est appliqué en fonction des intensités des ions provenant des pertes de neutre à partir de l'ion parent dans chacun des spectres MS/MS.

En effet, nous pouvons nous attendre à ce que, pour un peptide glycosylé donné, selon son état de glycosylation, les ions provenant de la (ou des) perte(s) de neutre(s) au sein du spectre aient une intensité variable.

Dans le cas d'un peptide monoglycosylé, l'unique ion provenant de la perte de neutre à partir de l'ion parent est relativement intense. Nous avons alors décidé de ne sélectionner que les spectres dont l'ion provenant de la perte de neutre à une intensité relative supérieure à 60% et un rang maximum au sein du spectre de 5.

Pour les spectres de peptides diglycosylés, l'intensité de la deuxième perte est moindre. Nous avons sélectionné les spectres dont l'ion le moins intense provenant des deux pertes de neutre à une intensité relative supérieure à 20% et un rang maximum au sein du spectre de 10.

Nous avons décidé, pour les peptides avec trois mannoses ou plus, de ne pas appliquer de filtres de ce type puisque, dans ce cas, la troisième perte de neutre peut avoir une intensité relativement faible.

III. Résultats et conclusion

Suite à l'analyse nanoLC-MS/MS des fractions peptidiques obtenues après digestion des protéines contenues dans chacune de ces 17 bandes, un grand nombre de queries a été obtenu, plus de deux fois supérieurs à ceux obtenus au cours de l'analyse des protéines contenus dans les surnageants de culture en milieu 7H9 (tableau 7).

	Sauton WT1	Sauton WT2
Nombre de spectres MS/MS enregistrés	236456	363529
Spectres avec signature de glycoconjugués (Neutral Loss)	8287	14513
% par rapport au nombre total de spectres MS/MS	3,50	3,99
Spectres de glycoconjugués putatifs de catégorie 3, avec interprétation de rang 1	3748	6964
Spectres de glycoconjugués putatifs avec attributions de séquences de peptides glycosylés	591	1103
Spectres de glycoconjugués putatifs avec un même nombre d'hexoses détectés par Mascot et par neutral Loss	338	827
Spectres de glycoconjugués avec un score d'attribution de séquence peptidique > 8	237	595
Spectres de glycoconjugués putatifs uniques avec un même nombre d'hexoses détectés par Mascot et par neutral Loss	125	292
Nombre de protéines glycosylées putatives uniques	117	283

Tableau 7 : Apport de l'outil "Neutral Loss" pour la selection des spectres de peptides glycosylés putatifs – milieu Sauton

Parmi ces spectres MS/MS, 3,5 à 4% sont sélectionnés par Neutral Loss comme spectres de glycoconjugués, contre 1,5% maximum dans l'étude précédente. La liste de spectres MS/MS obtenue en sortie de Neutral Loss a ensuite été interrogée par Mascot, et les différents filtres ont été appliqués : 237 spectres MS/MS pour le premier réplicat et 595 pour le deuxième sont validés comme spectres MS/MS de glycopeptides putatifs.

Nous nous sommes ensuite plus particulièrement intéressés aux protéines identifiées avec au moins deux peptides glycosylés. En effet, un plus grand risque de faux positif concerne les glycoprotéines identifiées avec un seul glycopeptide.

Dix glycoprotéines ont ainsi été identifiées et validées, dont deux qui l'ont été dans les deux réplicats : la β 1,3-glucanase et la lipoprotéine LprA (tableau 8). La lipoprotéine LprA est une des deux glycoprotéines validées avec deux peptides glycosylés dans l'étude précédente. De plus, les glycopeptides de huit des dix glycoprotéines identifiées ont de bonnes probabilités de glycosylation, selon NetOGlyc. Parmi ces 10 glycoprotéines, 6 sont des lipoprotéines. Or il a été décrit dans la littérature que les lipoprotéines ont de grandes probabilités de glycosylation (Herrmann, Delahay et al. 2000, Sutcliffe and Harrington 2004).

Les trois glycoprotéines connues n'ont pas été identifiées dans cette étude. Nous les avons alors recherchées dans la liste des interprétations peptidiques de Mascot afin d'évaluer une éventuelle perte de données sur ces glycoprotéines au cours de l'application des différents

filtres. En effet, les filtres appliqués sont relativement stringents et leur application pourrait entraîner la perte d'information sur des protéines glycosylées vraies positives.

En ce qui concerne la protéine Apa (Alanine and Proline rich Antigen), celle-ci est retrouvée dans la liste des protéines identifiées après interrogation Mascot de la liste de glycoconjugués putatifs de Neutral Loss, mais aucun des peptides identifiés n'est un peptide glycosylé.

Dans le cas du premier réplicat Sauton WT1, un peptide tri-glycosylé est identifié pour la protéine MPT83 et passe tous les filtres arbitrairement définis. En revanche, étant donné qu'un seul peptide est identifié pour cette protéine, celle-ci n'a pas été retenue dans la première liste de glycoprotéines validées. Le glycopeptide identifié contient les deux thréonines (Th48-Th49) identifiées comme sites de glycosylation. Dans le cas du deuxième réplicat, un peptide monoglycosylé de la protéine MPT83 est identifié, mais le score attribué à l'interprétation peptidique de ce peptide est inférieur à 8 et ne correspond alors pas aux critères fixés. De plus, ce peptide ne contient pas les deux thréonines sièges de la glycosylation, il semblerait alors que les filtres choisis permettent d'éliminer ce peptide faux-positif.

Enfin, après analyse des protéines du surnageant du premier réplicat, un peptide glycosylé est identifié pour la protéine SodC mais le nombre d'hexoses détectés sur le peptide est différent de celui détecté par Mascot et est alors éliminé du crible.

Afin de vérifier que les glycopeptides identifiés ne sont pas des faux-positifs, nous avons décidé d'analyser les protéines sécrétées par la souche mutante $\Delta Rv1002c$.

En effet, dans le cas où un peptide glycosylé identifié dans un des deux réplicats de la souche sauvage serait également identifié au cours de la recherche des glycoprotéines sécrétées de la souche mutante, ce peptide serait alors considéré comme aussi faux-positif. En effet, un spectre MS/MS correspondant à un peptide glycosylé faux-positif est un spectre dans lequel le ou les ion(s) identifié comme perte(s) de neutre par Mascot sont mal identifiés et correspondent en réalité à des ions y ou b de fragmentation du peptide. Pour une protéine identifiée à la fois dans les surnageants de culture de la souche sauvage et de la souche mutante (en milieu Sauton), la même erreur d'interprétation par Mascot devrait avoir lieu au cours de l'analyse des protéines de la souche mutante.

De plus, nous avons également décidé d'analyser les protéines du surnageant de culture de la souche complémentée, afin de confronter la liste des glycoprotéines identifiées après analyse des protéines de cette souche complémentée avec la liste des glycoprotéines identifiées dans les deux réplicats de la souche sauvage.

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

Gene Rv	Protéine	Expérience		AA	PTM	Degré glyco	Queries
Rv0315	Putative beta-1,3-glucanase	Sauton WT1	APAGPT(0,92)PAPAAPAAATGLLFHDEFDGPAGSVDPDSK	A		1	11
			PAGPT(0,92)PAPAAPAAATGLLFHDEFDGPAGSVDPDSK			1	6
		Sauton WT2	APAGPT(0,92)PAPAAPAAATGLLFHDEFDGPAGSVDPDSK			1	1
			PAGPT(0,92)PAPAAPAAATGLLFHDEFDGPAGSVDPDSK			1	1
Rv0838	Lipoprotéine LpqR	Sauton WT1	T(0,95)T(0,95)T(0,94)PS(0,96)GPVPPVS(0,94)EAAR	N		2	5
			T(0,95)T(0,95)T(0,94)PS(0,96)GPVPPVS(0,94)EAAR			3	1
			NT(0,95)T(0,95)T(0,94)PS(0,96)GPVPPVS(0,94)EAAR	V		2	1
Rv0928	Lipoprotéine PSTS3	Sauton WT2	GNDDNVTGGGATTGQASAK	C		2	2
						3	3
			DDNVTGGGATTGQASAK	N		2	1
Rv1270c	Lipoprotein putative LprA	Sauton WT1	AS(0,81)DT(0,77)AAT(0,69)ASNGDAAMLLK	K	D (N)	1	5
					D (N); O (M)	1	13
					D (N)	2	2
					D (N); O (M)	2	5
					D (N)	3	2
		Sauton WT2	AS(0,81)DT(0,77)AAT(0,69)ASNGDAAMLLK	K	D (N)	1	8
					D (N); O (M)	1	23
					D (N)	2	4
					D (N); O (M)	2	1
Rv1368	Lipoprotéine LprF	Sauton WT1	KPT(0,92)T(0,90)AS(0,58)S(0,59)PS(0,57)PGS(0,57)PS(0,58)PEAQQILQDSSK	K		1	2
						2	1
Rv1411c	Lipoprotéine LprG	Sauton WT1	VQVTKPPVS	Q		1	4
			TLSGDLTTNPTAATGNVK	K		1	1
Rv2068c	Beta lactamase	Sauton WT1	PAS(0,87)T(0,91)T(0,77)LPAGADLADR	R		2	1
			S(0,87)T(0,91)T(0,77)LPAGADLADR	A		2	1
						3	1
			T(0,91)T(0,77)LPAGADLADR	S		2	1
Rv2290	Lipoprotéine LppO	Sauton WT2	T(0,51)AT(0,83)PS(0,53)ES(0,56)GT(0,69)QT(0,68)T(0,66)R	R		1	1
						2	1

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

Rv2799	Protéine membranaire	Sauton WT1	SPIVAT(0,73)T(0,84)DPS(0,53)PFDPCR	K	C (C)	1	2
					C (C)	2	6
Rv3763	Lipoprotéine LpqH	Sauton WT2	GET(0,94)T(0,99)T(0,98)AAGT(0,95)T(0,96)AS(0,93)PGAAS(0,53)GPK	S		4	2
			S(0,80)T(0,93)T(0,83)GET(0,94)T(0,99)T(0,98)AAGT(0,95)T(0,96)AS(0,93)PGAAS(0,53)GPK	K		2	1
						3	1
						4	1

Tableau 8: Liste des glycoprotéines identifiées après analyse des deux réplicats de la souche sauvage, milieu Sauton. En bleu, probabilité de glycosylation des sérines et thréonines selon le logiciel NetOGlyc.

E. Recherche des glycoprotéines sécrétées de *Mycobacterium tuberculosis* – Milieu de culture Sauton, souches mutante et complémentée

I. Echantillons biologiques, conditions analytiques et retraitement des données

De la même façon qu’au cours de l’étude précédente, les surnageants de culture des deux souches mutante et complémentée ont été recueillis et filtrés à un temps de culture de six semaines, et les fractions peptidiques ont été obtenues de la même façon que précédemment.

De plus, le workflow de retraitement des données est le même que celui utilisé précédemment (figures 36 et 39).

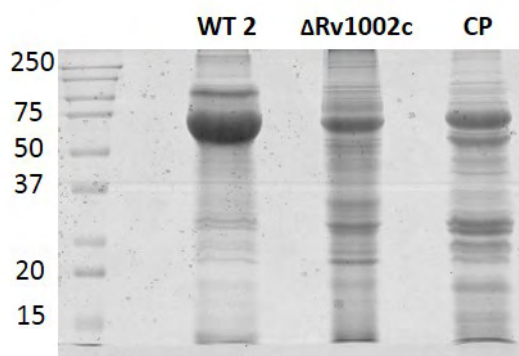


Figure 40: Gel SDS-Page réalisé avec les protéines des surnageants de la souche sauvage, de la souche mutante et de la souche complémentée.

En ce qui concerne le gel SDS-Page des protéines de la souche mutante, la bande correspondant à la protéine de type HSP possède une intensité fortement diminuée (figure 40). De plus, les protéines sont réparties de façon plus homogène sur la piste de gel, par rapport à la piste correspondant au deuxième réplicat de la souche sauvage par exemple.

Cette bande est également moins intense sur la piste correspondant aux protéines de la souche complémentée, alors des bandes à des poids moléculaires plus faibles deviennent plus intenses. Il semblerait que les profils protéiques des deux souches sont différents, nous verrons alors si la glycosylation des protéines diffère entre la souche sauvage et cette souche complémentée.

1. Analyse des protéines de la souche mutante

L'homogénéité dans la répartition des protéines sur la piste de gel peut expliquer le nombre de queries plus important obtenu après analyse des peptides de cette souche, par rapport aux deux réplicats biologiques de la souche sauvage (tableau 9, 100 000 queries supplémentaires).

De façon surprenante, le pourcentage de spectres de glycoconjugués putatifs par rapport au nombre total de queries est similaire à ceux obtenus après analyse des protéines des souches sauvages. De plus, après application de chacun des filtres, 310 spectres MS/MS de glycopeptides putatifs restent validés. En revanche, moins de 5% de ces spectres sont des spectres de peptides multiglycosylés putatifs. Enfin, parmi ces 310 spectres, aucun ne correspond à la même protéine, et aucune protéine glycosylée n'est alors validée.

Cette étude montre alors l'efficacité du workflow adopté puisque l'absence de glycosylation des protéines de la souche mutante $\Delta Rv1002c$ est bien confirmée.

	Sauton WT1	Sauton WT2	Sauton $\Delta 1002c$	Sauton CP
Nombre de spectres MS/MS enregistrés	236456	363529	456736	380336
Spectres avec signature de glycoconjugués (Neutral Loss)	8287	14513	14287	19575
% par rapport au nombre de spectres MS/MS total	3,5	3,99	3,13	5,15
Spectres de glycoconjugués putatifs de catégorie 3, avec interprétation de rang 1	3748	6964	8045	9654
Spectres de glycoconjugués putatifs avec un même nombre d'hexoses détectés par Mascot et par « Neutral Loss »	338	827	648	1514
Spectres de glycoconjugués avec un score d'attribution de séquence peptidique > 8	237	595	399	1161
Spectres de glycoconjugués putatifs uniques avec un même nombre d'hexoses détectés par Mascot et par neutral Loss	125	292	218	514
Nombre de protéines glycosylées putatives uniques	117	283	215	481

Tableau 9: Apport de l'outil "Neutral Loss" pour la sélection des spectres de peptides glycosylés putatifs – milieu Sauton, souches sauvage, mutante et complémentée

2. Analyse des protéines de la souche complémentée

En ce qui concerne la souche complémentée, 5.15% du total des spectres sont des spectres de glycoconjugués putatifs sélectionnés par Neutral Loss (tableau 9). Après retraitement des

données et application de chacun des filtres, ce sont plus de 1000 spectres MS/MS de glycopeptides putatifs qui sont validés. Finalement, ce sont 18 glycoprotéines qui sont validées avec au moins deux peptides glycosylés différents, dont 7 qui ont précédemment été validées dans au moins un des deux réplicats de la souche sauvage, ce qui permet d'augmenter la fiabilité de ces résultats (tableau 10).

De plus, dans cette étude, la glycoprotéine Apa est identifiée avec trois peptides glycosylés de la région C-terminale de la protéine, région dans laquelle les sites de glycosylation de la protéine ont été définis (Dobos, Khoo et al. 1996).

La lipoprotéine LpqH est également ici identifiée avec deux glycopeptides différents, et un nombre d'hexoses allant jusqu'à 8. Les sites de glycosylation de cette protéine décrits dans la littérature sont contenus dans le glycopeptide identifié dans cette étude (Herrmann, O'Gaora et al. 1996).

La protéine Apa, la beta-lactamase (Rv2068c), trois protéines membranaires (Rv2799, Rv3785c, Rv3835) et la β -1,3-glucanase (Rv0315) ont été décrites comme parmi les plus abondantes des protéines des surnageants de culture, il est alors cohérent de les retrouver dans cette étude (de Souza and Wiker 2011).

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

Gene Rv	Protéines	Expérience		Peptide glycosylé	PTM	PTM	Degré glyco	Queries
0175	MCE-associated membrane protein	Sauton CP	K	DCVAATQAPDAGAMSASMQK	C (C)		1	10
			R	VKVSNTDSDAH	D (N)		1	1
0822c	LytR/CpsA/Psr family protein	Sauton CP	R	LGNTPSTPPTTK			1	31
							2	1
0838	Probable conserved lipoprotein lpqR	Sauton CP	V	NT(0.95)T(0.95)T(0.94)PS(0.96)GPVPPVS(0.94)EAAR	D (N)		2	5
							2	4
							3	3
					D(N)		3	1
							4	6
			N	T(0.95)T(0.95)T(0.94)PS(0.96)GPVPPVS(0.94)EAAR			2	6
							3	6
							4	6
							5	1
1270c	Putative lipoprotein LPRA	Sauton CP	K	AS(0.81)DT(0.77)AAT(0.69)ASNGDAAMLLK	D (N)		1	38
					D (N)	O (M)	1	18
					D (N)		2	7
					D (N)	O (M)	2	15
1368	LprF	Sauton CP	A	S(0.58)S(0.59)PS(0.57)PGS(0.57)PS(0.58)PEAQQLQDSSK			1	6
			T	AS(0.58)S(0.59)PS(0.57)PGS(0.57)PS(0.58)PEAQQLQDSSK			3	1
1754c	Predicted protein	Sauton CP	R	ANPGNPAPNPAT(0.91)GPQPPNAT(0.93)QLT(0.96)PNPAPAPDPAPA			1	3
				TSRWGRT(0.59)SSPT(0.59)T			3	1
1860	Apa	Sauton CP	A	GEVAPT(0.94)PT(0.65)T(0.91)PT(0.87)PQR			1	9
							2	1
							3	11
			R	PLVAPPPAPAPAPAEAPAPAPAGEVAPT(0.94)PT(0.65)T(0.91)PT(0.87)PQR			1	2
			L	VAPPPAPAPAPAEAPAPAPAGEVAPT(0.94)PT(0.65)T(0.91)PT(0.87)PQR			1	34
2068c	Beta-lactamase	Sauton CP	P	AS(0.87)T(0.91)T(0.77)LPAGADLADR			2	8
			A	S(0.87)T(0.91)T(0.77)LPAGADLADR			2	3
2290	LppO	Sauton CP	R	T(0.51)AT(0.83)PS(0.53)ES(0.56)GT(0.69)QT(0.68)T(0.66)R			1	12
			R				2	1

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

2394	Putative gamma-glutamyltranspeptidase GGTB	Sauton CP	A	GAP S(0.78)T(0.56) AGPCEIVPNG T(0.93) PAPK	C (C)	D (N)	1	3
			A	GPCEIVPNG T(0.93) PAPK	C (C)	D (N)	1	8
2799	Probable membrane protein	Sauton CP	V	AT (0.73)T(0.84) DPSPFDPK	C (C)		1	12
			K	SPIVAT (0.73)T(0.84)DPS(0.53) PFDPK	C (C)		1	7
			K		C (C)		2	23
			K		C (C)		3	2
			A	T(0.73)T(0.84)DPS(0.53) PFDPK	C (C)		1	8
			A		C (C)		2	1
			I	VAT (0.73)T(0.84)DPS(0.53) PFDPK	C (C)		1	1
3244c	Lipoprotein lpqB	Sauton CP	I	GT (0.76) VERPV S(0.94) NLPK S(0.79) PGMDPDVLLR			2	2
							3	1
3491	Putative uncharacterized protein	Sauton CP	K	QPFSLQLIGPPSPVQR			1	5
			F	SLQLIGPPSPVQR			1	12
3587c	Putative conserved membrane protein	Sauton CP	R	SLPVPFILNQPPPPGPVPAPGPAQAPP S(0.66) PAQGG			1	1
			A	APPQGQNP T(0.99)PT(0.99)PT(0.98) AAVQPPVVK			3	1
			Q	GQNP T(0.99)PT(0.99)PT(0.98) AAVQPPVVK			3	3
3705a	Conserved hypothetical proline rich protein	Sauton CP	A	VGPTGPG T(0.62) TAPARP			2	13
							3	1
3706c	Conserved hypothetical proline rich protein	Sauton CP	A	AVRPGPGGGPGQVP S(0.78)S(0.92)VS(0.76)PPAT(0.67) PAP			2	9
							3	7
				VRPGPGGGPGQVP S(0.78)S(0.92)VS(0.76)PPAT(0.67) PAP			2	9
							3	6
							4	1
3763	19 kDa lipoprotein antigen LPQH	Sauton CP	A	AG T(0.95)T(0.96)AS(0.93)PGAAS(0.53) GPK			1	1
			K	S(0.80)T(0.93)T(0.83)GS(0.94)GET(0.94)T(0.99)T(0.98)AAGT(0.95)T(0.96)AS(0.93)PGAAS(0.53) GPK			3	2
							4	11
							5	8
							6	9
							7	4
							8	4
3835	Putative conserved membrane protein	Sauton CP	R	LNLPIPLQL T(0.98) PR			1	12
			R	TFPGMEYGQNAAPPSPAR			1	5

Tableau 10: Liste des glycoprotéines identifiées après analyse des protéines du surnageant de la souche complétée de *Mycobacterium tuberculosis* en milieu Sauton. En bleu, probabilité de glycosylation des sérines et thréonines selon le logiciel NetOGlyc.

F. Protéines cellulaires

Afin d'obtenir les extraits de protéines cellulaires, les bactéries ont subies une lyse mécanique.

Par la suite, 100µg de protéines issues de la lyse des bactéries de la souche sauvage et 100µg de protéines issues de la lyse des bactéries de la souche complémentée ont été déposées sur un gel SDS-Page avec fractionnement (figure 41).

Les profils des deux pistes de ce gel sont similaires avec des protéines réparties de façon uniformes sur les pistes.

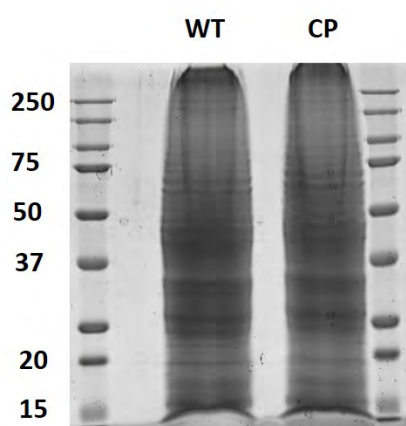


Figure 41: Gel SDS-Page réalisé avec les protéines issues de la lyse des bactéries de la souche sauvage et de la souche complémentée.

Une digestion *in gel* des protéines a été réalisée et, après analyse par nanoLC-MS/MS des peptides, plus de 4000 protéines ont été validées. 4012 protéines de *Mycobacterium tuberculosis* sont annotées (Schubert, Mouritsen et al. 2013), nous avons donc eu dans cette étude une vue exhaustive des protéines cellulaires de la bactérie.

Après application du workflow de retraitement des données pour la recherche des glycopeptides, nous avons identifié et validé 10 glycoprotéines (tableau 11), dont 4 précédemment identifiées lors de l'analyse des protéines sécrétées des souches sauvage et complémentée du milieu Sauton : les lipoprotéines LprF, LpqH, LprA et LprG. Pour ces 4 glycoprotéines putatives, les glycopeptides identifiés sont identiques à ceux identifiés précédemment. Pour les 6 autres glycoprotéines nouvellement validées, 3 ont été identifiées avec des peptides possédant de bonnes probabilités de glycosylation.

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

Gene Rv	Protéines	Expérience	AA	Peptide glycosylé	PTM	PTM	Degré glyco	Queries
0928	Periplasmic phosphate-binding lipoprotein PSTS3	EC CP	C	GNDDNVTGGGATTGQASAK			2	2
			N	DDNVTGGGATTGQASAK			3	3
1270c	Putative lipoprotein LPRA	EC CP	K	AS(0.81)DT(0.77)AAT(0.69)ASNGDAAMLLK	D (N)	O (M)	1	6
					D (N)		1	7
					O (M)		1	1
							1	2
					D (N)		2	5
							2	2
					D (N)	O (M)	2	3
					O (M)		2	1
					D (N)	O (M)	3	2
					D (N)		3	1
							3	1
					D (N)		4	2
		EC WT	K	AS(0.81)DT(0.77)AAT(0.69)ASNGDAAMLLK	D (N)		2	2
					D (N)	O (M)	2	1
					D (N)		1	2
					D (N)	O (M)	1	1
1368	LprF	EC CP	K	KPT(0.921)T(0.902)AS(0.575)S(0.589)PS(0.571)PGS(0.569)PS(0.580)PEAQQLQDSSK			1	3
							2	3
							3	2
							4	1
		EC WT					4	1
							3	2
							2	2
							1	2
1411c	Putative conserved lipoprotein LPRG	EC CP	Q	VQVTKPPVS			1	5
			K	TLSGDLTTNPTAATGNVK			1	2
		EC WT	Q	VQVTKPPVS			1	1
			K	TLSGDLTTNPTAATGNVK			1	1

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

1743	Serine/threonine-protein kinase pknE	EC CP	R	T(0.72)S(0.89)APPPPPADAVELR			3	1
							2	6
2969c	Possible conserved membrane or secreted protein	EC CP	K	VPDCINSGKYEKVDGLAAAVNVH	D (N)	P (C)	1	1
			K	IKEIVGDVPGIDSAATA			1	1
			K	IKEIVGDVPGIDSAATATS			1	2
		EC WT	K	IKEIVGDVPGIDSAATATS			1	2
			K	VDGLAAAVNVHATPTVR			1	1
3584	Lipoprotein lpqE	EC CP	K	AGQGSVMVPISAGLATPHE			1	1
					O (M)		1	1
		EC WT	K	AGQGSVMVPISAGLATPHE			1	1
					O (M)		1	1
3623	Lipoprotein lpqG	EC CP	K	VISISEASGAAPT(0.6)HEAPAPPR			1	3
			R	VT(0.76)VAPQYSNPEAGTATITGYR			1	1
3763	19 kDa lipoprotein antigen LPQH	EC CP	K	S(0.80)T(0.93)T(0.83)GS(0.94)GET(0.94)T(0.99)T(0.98)AAGT(0.95)T(0.96)AS(0.93)PGAAS(0.53)GPK			1	1
							2	2
							3	2
							4	2
							6	1
							7	1
							8	1
3835	Putative uncharacterized protein	EC CP	R	LAGYIAS(0.66)NPVP(0.76)T(0.69)GAK			3	1
							2	1
							1	1

Tableau 11: Liste des glycoprotéines identifiées après analyse des protéines du surnageant de la souche complémentée en milieu Sauton

G. Apport de l'outil « Neutral Loss » par rapport à une stratégie de recherche classique des glycopeptides avec le moteur de recherche Mascot

I. Confrontation de deux approches

Une fois la stratégie de recherche des glycoprotéines mise au point, nous avons essayé d'évaluer et de quantifier l'apport de l'utilisation de l'outil « Neutral Loss » par rapport à une stratégie classique de recherche des protéines avec le logiciel Mascot. Nous avons comparé deux approches (figure 42), en travaillant sur les données acquises après analyse des protéines de la souche complémentée (pour laquelle nous avons le plus de glycopeptides putatifs) :

- **Première approche : qu'apporte l'utilisation seule du moteur de recherche Mascot pour la recherche de glycoprotéines ?**

Les attributions peptidiques de chacun des spectres MS/MS acquis au cours de l'expérience, après interrogation des banques protéiques par Mascot, sont récupérées. Nous avons regardé quelles sont les informations relatives aux peptides glycosylés que peut apporter cette approche, sans utilisation de « Neutral Loss ».

- **Deuxième approche : qu'apporte l'utilisation combinée du moteur de recherche Mascot et de « Neutral Loss » pour la recherche de glycoprotéines ?**

D'une part, « Neutral Loss » sélectionne les spectres MS/MS de glycoconjugués putatifs parmi la liste de tous les spectres MS/MS acquis au cours de l'analyse de chacune des fractions peptidiques. D'autre part, les attributions peptidiques, après interrogation Mascot de chacun des spectres MS/MS acquis au cours de l'expérience, sont recueillies. Ensuite, pour chacun des spectres MS/MS retenus par « Neutral Loss » comme spectres MS/MS de glycoconjugués putatifs, l'interprétation de séquence peptidique de Mascot est apposée.

Contrairement au workflow utilisé jusque-là, ce n'est pas la peak list réduite de sortie de « Neutral Loss » qui est soumise à l'interrogation Mascot.

II. Résultats

Tout d'abord, pour l'approche n°1, en amont du workflow, la liste de spectres MS/MS de départ est de 343257 spectres (tous les spectres enregistrés au cours de l'analyse des 17

digestats peptidiques), et de 18251 pour l'approche n°2 (tous les spectres enregistrés au cours de l'analyse des 17 digestats peptidiques et sélectionnés par « Neutral Loss »).

Avec l'approche n°1, seulement 2.56% des spectres ont une attribution de séquence de peptide glycosylé, contre 29% avec l'approche n°2. L'utilisation de Neutral Loss, couplée à l'utilisation de Mascot semble alors permettre de pré-enrichir la « peak list » de départ en peptides glycosylés putatifs.

Ensuite, afin d'évaluer la qualité des interprétations peptidiques des peptides glycosylés putatifs dans les deux approches, nous avons regardé les scores attribués à ces spectres. Avec l'approche n°1, 56% des spectres ont un score d'attribution peptidique supérieur à 8 contre 58% avec l'approche n°2.

Avec l'approche n°2, nous avons ensuite appliqué le filtre n°3 (figure 39) sur nombre de pertes de neutre détecté par Mascot identique à celui détecté par « Neutral Loss ». Après application de ce filtre, 62% des 2691 spectres sélectionnés possèdent un score d'attribution peptidique supérieur à 8. Nous avons ensuite appliqué le filtre n°5 (figure 39) sur les intensité de perte(s) de neutre au sein des spectres, et, finalement, 78% des 876 spectres sélectionnés possèdent alors un score d'attribution peptidique supérieur à 8. L'application de ces deux filtres semble alors augmenter de façon significative la qualité des spectres de peptides glycosylés putatifs.

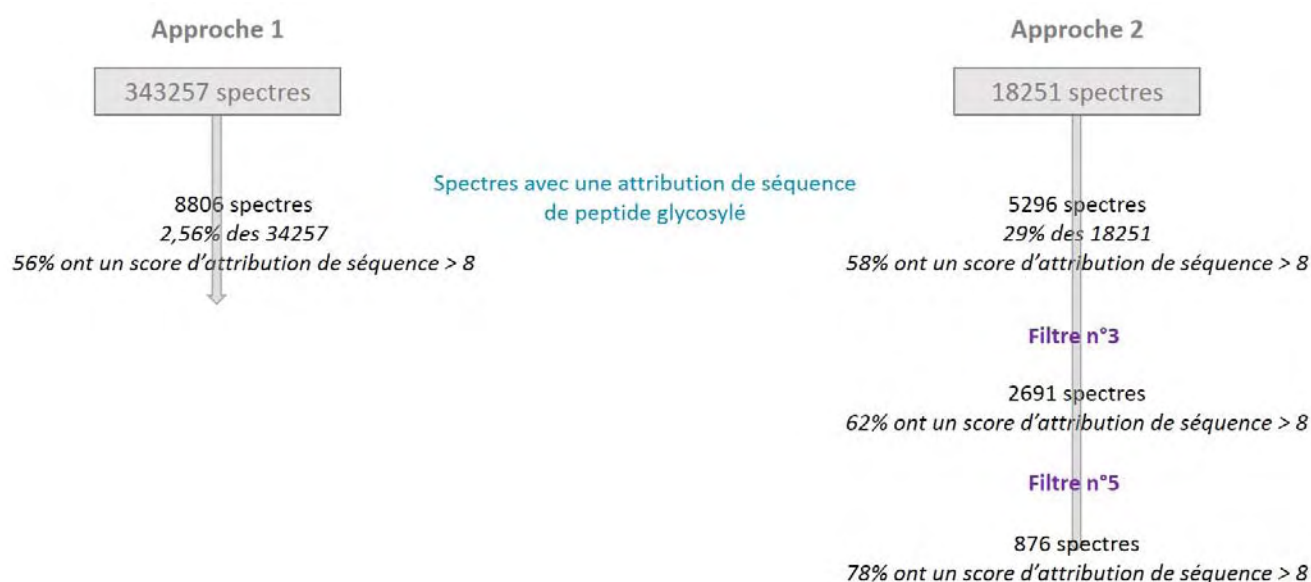


Figure 42: Workflow pour la comparaison d'une stratégie de recherche des glycoprotéines avec l'utilisation seule du moteur de recherche Mascot par rapport à l'utilisation combinée de « Neutral Loss » et de Mascot

Certains spectres attribués à des peptides glycosylés selon Mascot ne sont pas sélectionnés par Neutral Loss. Afin de savoir si ces spectres correspondent à des faux-positifs ou à des vrais-positifs dont on perdrait l'information avec Neutral Loss, nous nous sommes intéressés à ces spectres, au nombre de 77 dans le premier des 17 digestats peptidiques analysés (première bande découpée du gel). En effet, dans ce digestat, 146 spectres ont une attribution de peptide glycosylé selon Mascot et 69 sont sélectionnés par « Neutral Loss ». Le score moyen d'attribution peptidiques de ces spectres est relativement faible (7.36), et 50% des attributions sont des attributions de séquences avec des clivages manqués. Dix de ces spectres ont été analysés de façon plus précise. Pour les 9 spectres de peptides putatifs monoglycosylés, l'intensité relative au sein des spectres de l'ion provenant de la perte de neutre à partir de l'ion parent est inférieure à 9%, et le rang est supérieur à 20. Pour le spectre du peptide putatif diglycosylé, l'intensité relative de l'ion le moins intense des deux ions provenant des pertes de neutre est de 4% et le rang est supérieur à 20. Ces spectres sont alors des spectres de mauvaise qualité, et l'utilisation de « Neutral Loss » permet de les éliminer lors d'une recherche des glycopeptides.

H. Conclusions & perspectives

Nous avons montré que l'outil bioinformatique « Neutral Loss » est un outil performant pour la recherche de spectres MS/MS de glycopeptides putatifs parmi des jeux de données relativement importants. Son utilisation, couplée à l'utilisation du moteur de recherche Mascot pour l'obtention des attributions de séquences peptidiques pour chacun des spectres MS/MS, permet d'obtenir une liste de glycopeptides putatifs avec un taux de confiance important.

Après analyse de chacune des conditions étudiées (protéines des surnageants de culture d'un milieu 7H9, avec ou sans méthode d'égalisation des protéines, ainsi que d'un milieu Sauton, et protéines cellulaires de bactéries issues d'une culture en milieu Sauton), nous avons identifié 25 glycoprotéines, sans utiliser aucune méthode d'enrichissement (tableau 12). Parmi ces protéines, les protéines Apa et LpqH, bien caractérisées, sont retrouvées. De plus, la protéine codée par le gène Rv0175 est une protéine associée au complexe mce. Ce complexe a un rôle dans l'activation des gènes impliqués dans le trafic des substrats pendant la phase initiale d'infection (Stavrum, Valvatne et al. 2012). La protéine pknE, également identifiée comme glycoprotéine a été montrée comme ayant un rôle dans la survie de *Mycobacterium tuberculosis*, en inhibant l'apoptose des macrophages infectés (Jayakumar, Jacobs et al. 2008). Nous avons alors ici mis en évidence la glycosylation de protéines impliquées dans la virulence du bacille.

RESULTATS première partie : Analyse du glycoprotéome de *Mycobacterium tuberculosis*

Gene Rv	Protéines	Expérience
0175	MCE-associated membrane protein	SN CP
0315	Putative beta-1,3-glucanase	SN WT1
		SN WT2
0822c	LytR/CpsA/Psr family protein	SN CP
0838	Probable conserved lipoprotein lpqR	SN CP
		SN WT1
0928	Periplasmic phosphate-binding lipoprotein PSTS3	SN WT2
1270c	Putative lipoprotein LPRA	SN ADC PMa
		SN CP
		SN DC
		SN DC PMa
		SN DC PMn
		SN WT1
		SN WT2
		EC CP
1368	LprF	EC WT
		SN CP
		SN WT1
		EC CP
1411c	Putative conserved lipoprotein LPRG	EC WT
		SN WT1
		EC CP
1743	Serine/threonine-protein kinase pknE	EC CP
1754c	Predicted protein	SN CP
1860	Apa	SN CP
2068c	Beta-lactamase	SN CP
		SN DC
		SN WT1

Gene Rv	Protéines	Expérience
2290	LppO	SN CP
		SN WT2
2394	Putative gamma-glutamyltranspeptidase GGTB	SN CP
2799	Probable membrane protein	SN ADC
		SN ADC PMn
		SN CP
		SN DC
		SN WT1
2969c	Possible conserved membrane or secreted protein	EC CP
		EC WT
3244c	Lipoprotein lpqB	SN CP
3491	Putative uncharacterized protein	SN ADC
		SN ADC PMn
		SN CP
		SN DC PMn
3584	Lipoprotein lpqE	EC CP
		EC WT
3587c	Putative conserved membrane protein	SN CP
3623	Lipoprotein lpqG	EC CP
		EC WT
3705a	Conserved hypothetical proline rich protein	SN CP
3706c	Conserved hypothetical proline rich protein	SN CP
3763	19 kDa lipoprotein antigen LPQH	SN CP
		SN WT2
		EC CP
3835	Putative uncharacterized protein	EC CP
		SN CP

Tableau 12: Liste des glycoprotéines identifiées avec le mode de retraitement des données décrit dans les figures 34 et 37. Les données issues de l'analyse des protéines du surnageant d'une culture en milieu 7H9 ont été reprises avec ce mode de retraitement.

Dans cette étude, nous avons évalué l'apport de deux méthodes d'enrichissement des glycoprotéines ou glycopeptides : l'affinité à la lectine concanavaleine A et la chromatographie d'interaction hydrophile de type HILIC. Ces deux méthodes ne nous ont pas permis d'obtenir des résultats satisfaisants et/ou reproductibles, c'est pourquoi nous avons décidé de ne pas les utiliser pour cette recherche des glycoprotéines de *Mycobacterium tuberculosis*.

De plus, utiliser de telles méthodes d'enrichissement en amont de l'analyse des protéines ou peptides par spectrométrie de masse peut conduire à analyser seulement une sous-population des glycoprotéines présentes dans l'échantillon. En effet, par exemple, l'affinité de la concanavaleine A pour les motifs mannosylés est une affinité faible, et certaines glycoprotéines peuvent ne pas se lier à celle-ci et ne pas être analysées au cours d'une étude utilisant cette méthode d'enrichissement. De plus, cette méthode n'est pas spécifique des

glycoprotéines ou glycopeptides et dans le cas d'échantillons contenant également d'autres glycoconjugués, ceux-ci peuvent saturer les sites de fixation de la lectine qui ne peut alors plus fixer les glycopeptides ou glycoprotéines d'intérêt. Enfin, l'utilisation d'une méthode de recherche des glycoprotéines avec utilisation d'une méthode d'enrichissement entraîne une variabilité entre les différents échantillons à analyser et éventuellement à comparer.

La méthode de recherche des glycoprotéines que nous avons utilisée permet de s'affranchir de ces problèmes puisqu'elle consiste à filtrer toutes les données acquises au cours d'une expérience protéomique de type nanoLC-MS/MS. Au cours de ce type d'expérience de glycoprotéomique, la seule variabilité entre les analyses faites est celle liée à la préparation des échantillons et à l'analyse des digestats en nanoLC-MS/MS.

De plus, l'outil « Neutral Loss » est un outil dont l'utilisation est rapide puisque, avec un set de données de plus de 300 000 spectres, la liste de spectres de glycoconjugués putatifs est obtenue en moins d'une heure.

Une des limites de cette stratégie de glycoprotéomique est une des limites de toute stratégie de protéomique : certains peptides sont non « MS observables » (Kuster, Schirle et al. 2005). En effet, la plage de masse pour la détection des peptides dans l'Orbitrap est de 300 à 2000 m/z. Du fait que les motifs oligosaccharidiques portés par les peptides accroissent leurs masses, certains des peptides obtenus peuvent posséder une masse au-delà de 2000 m/z. Pour éviter ce problème, nous avons décidé de réaliser une double digestion des protéines, avec la trypsine et l'endoprotéinase GluC-V8. Cependant, cette double digestion ne nous a permis d'identifier que trois glycoprotéines : Apa, LprG ainsi qu'une protéine putative non identifiée jusque là. Il serait alors nécessaire de répéter cette expérience en améliorant le protocole de préparation des échantillons ainsi que le retraitement des données pour obtenir une liste plus complète.

Afin de mieux comprendre le phénotype de virulence atténué pour la souche mutante $\Delta Rv1002c$ du bacille, nous avons par la suite entrepris de réaliser une analyse protéomique quantitative des protéines de la souche mutante, de la souche sauvage et de la souche complémentée, pour essayer de mettre en évidence des protéines différentiellement exprimées entre ces 3 souches et des voies métaboliques affectées par un arrêt du phénomène de glycosylation.

RESULTATS deuxième partie: Recherche des voies métaboliques de *Mycobacterium tuberculosis* impactées par un arrêt du processus de glycosylation

Plusieurs méthodes d'analyse quantitative des données de nanoLC-MS/MS sont disponibles et peuvent être classées en deux grandes catégories: les méthodes de quantification absolue et les méthodes de quantification relative, parmi lesquelles la méthode de quantification sans marquage « label-free » que nous avons utilisée dans cette étude.

A. Les méthodes de quantification relative

Les méthodes de quantification relative sont utilisées pour quantifier et comparer des protéomes globaux, sans hypothèse de départ sur les espèces potentiellement variantes.

Une première approche a consisté à séparer des mélanges protéiques sur des gels d'électrophorèse bidimensionnelle (gels 2D) et à comparer l'intensité des spots sur les différents gels, correspondant alors à des protéines variantes (Patterson and Aebersold 2003).

Depuis, grâce à la montée en puissance des approches haut-débit d'identification des protéines et le développement d'outils bio-informatiques pour des approches comparatives, des méthodes de quantification relative basées sur l'interprétation des données acquises au cours d'expériences de nanoLC-MS/MS sont possibles. Pour cela, des méthodes avec marquage isotopique et des méthodes sans marquage ou « label free » peuvent être utilisées.

I. Les méthodes avec marquage isotopique

Au cours d'une méthode de quantification avec marquage isotopique, un des deux échantillons à comparer, l'échantillon « lourd » est par exemple marqué avec un isotope stable (^2H , ^{13}C , ^{15}N et ^{18}O), puis ajouté à l'échantillon « léger » non marqué. Les deux échantillons sont traités et analysés simultanément.

Le marquage peut se faire à différents niveaux (figure 43): par voie chimique au niveau des peptides (approche iTRAQ par exemple, (Thompson, Schäfer et al. 2003)) ou au niveau des protéines (approche ICAT par exemple (Gygi, Rist et al. 1999)), ou encore par voie métabolique au moment de la culture cellulaire par exemple, en amont du processus analytique (approche SILAC (Ong and Mann 2005)).

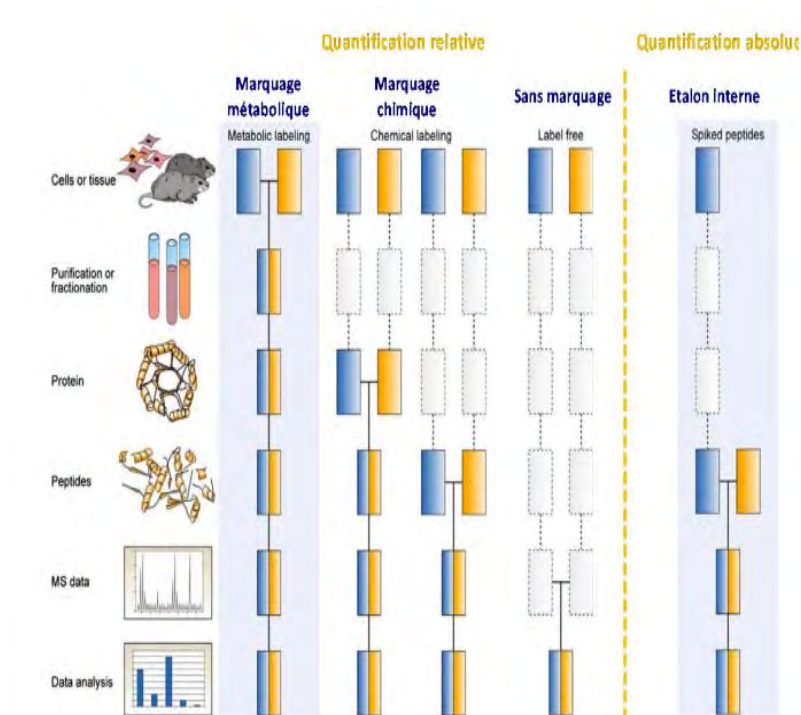


Figure 43: Les strategies de protéomique quantitative basées sur la spectrométrie de masse

Les échantillons à comparer constitués d'un échantillon « léger » et d'un échantillon « lourd » sont mélangés puis traités et analysés simultanément par LC-MS/MS. Un peptide léger et son peptide lourd homologue présentent les mêmes caractéristiques physico-chimiques et vont être élués au même moment de la colonne chromatographique. Ils ont également une faible différence de masse au sein des spectres de masse. En comparant l'intensité des deux peptides dans le spectre MS, une quantification relative est obtenue (Bantscheff, Lemeer et al. 2012).

II. Les méthodes sans marquage isotopique ou « label-free »

1. Quantification sans marquage par extraction du signal MS

Il est également possible d'obtenir une quantification relative de deux échantillons sans réaliser de marquage isotopique de l'un d'entre eux, on parle alors de méthode « label-free ». En effet, l'apparition de spectromètres de masse à très haute résolution (permettant d'obtenir une très grande précision de masse sur les ions peptidiques), la très grande vitesse de séquençage des analyseurs, ainsi que le développement d'outils bioinformatiques de plus en plus performant permettent aujourd'hui une quantification relative des protéines au sein de plusieurs échantillons en comparant l'intensité du signal MS mesuré pour un même

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

peptide dans plusieurs expériences, il s'agit alors d'une quantification « label-free » par extraction du signal MS (figure 44) (Mueller, Brusniak et al. 2008).

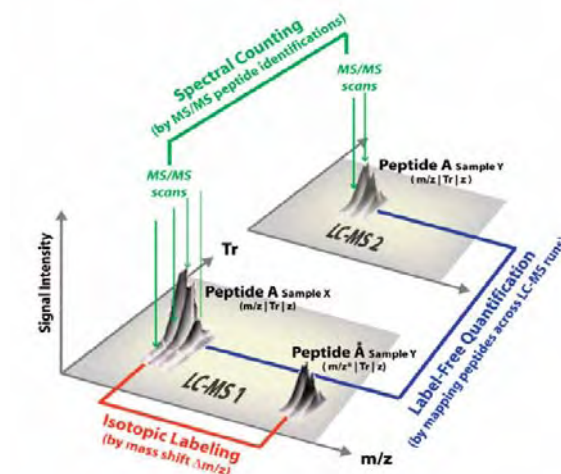


Figure 44: Vue d'ensemble de différentes approches de quantification par LC-MS.

Les approches avec marquage isotopique permettent la quantification de protéines au sein d'une même analyse alors qu'avec les méthodes sans marquage, la quantification est réalisée sur deux expériences différentes. En bleu, la méthode de quantification par extraction du signal MS. En vert, la méthode de quantification par « spectral counting », paragraphe suivant.

Cette méthode nécessite l'extraction de l'ensemble des signaux MS et leur appariement aux identifications peptidiques correspondantes. La haute résolution permet de limiter les faux appariements de signaux peptidiques entre les conditions. De plus, cette méthode permet de quantifier des protéines relativement peu abondantes avec une bonne précision.

Cependant, cette méthode nécessite plusieurs étapes bioinformatiques assez délicates, comme l'alignement en temps de rétention des différentes analyses à comparer, la normalisation du signal du spectromètre entre les analyses, ou l'extraction du signal entre les différentes analyses.

Même si cette méthode de quantification n'est pas aussi précise qu'une méthode avec marquage, cette approche est de plus en plus utilisée dans la communauté scientifique, en raison de l'amélioration constante des performances de spectromètres de masse en terme de haute résolution et de vitesse de séquençage, la robustesse croissante des système HPLC, et le développement d'outils bioinformatiques de plus en plus performants (Liu, Dekker et al. 2013).

2. L'analyse bio-informatique des données de protéomique quantitative sans marquage par extraction du signal MS

Les approches d'extraction du signal MS nécessitent l'utilisation de logiciels permettant d'extraire les intensités des signaux peptidiques et de les relier aux identifications correspondantes. Pour cela, de nombreux logiciels bio-informatiques ont été développés, comme Maxquant (Cox and Mann 2008) et MFPaQ (Bouyssie, Gonzalez de Peredo et al. 2007), développé au laboratoire.

L'extraction des signaux MS nécessite un alignement au préalable des temps de rétention de tous les peptides entre les runs des conditions à comparer, puis d'associer ces signaux MS aux identifications des peptides (données par le logiciel Mascot par exemple).

Deux approches sont possibles pour comparer les signaux MS de plusieurs conditions en « label-free ».

La première consiste à utiliser comme point de départ les résultats d'identification MS/MS. A partir de ces résultats, les signaux MS des peptides sont extraits sous forme d'un appel d'ion ou XIC (« Extracted ion chromatogram »). En effet, l'intensité du signal MS d'un peptide qui élue de la colonne chromatographique peut être suivie au cours du temps. L'aire sous la courbe du pic chromatographique est le courant ionique extrait (XIC) et est proportionnelle à l'abondance du peptide dans l'échantillon (Ong and Mann 2005). Cette méthode consiste donc à extraire l'intensité du signal pour chaque peptide identifié, de m/z connu.

La seconde approche est centrée sur l'établissement de cartes LC-MS. Les logiciels vont définir des empreintes MS des peptides sur la base de la modélisation théorique des massifs isotopiques peptidiques et des profils d'élution de ces peptides. La quantification relative est réalisée en alignant les cartes et en comparant les aires de ces empreintes peptidiques dans les différents échantillons. Ce n'est que dans un second temps que le lien avec les interprétations peptidiques est réalisé.

3. Quantification sans marquage par « spectral counting »

En 2004, Liu et al. (Liu, Sadygov et al. 2004) démontrent qu'il existe une forte corrélation entre le niveau d'abondance relative d'une protéine et le nombre de spectres MS/MS qui en sont issus. En effet, en mode DDA, les ions précurseurs les plus intenses sont sélectionnés puis fragmentés. Plus une protéine est abondante, plus les peptides qu'elle génère sont abondants et ces peptides sont souvent séquencés dans le spectromètre de masse. Ainsi, une quantification relative des protéines peut être effectuée en comparant le nombre de spectres MS/MS enregistrés pour une protéine donnée dans plusieurs séries d'expériences. On parle alors de quantification « label-free » par « spectral counting ».

Cette méthode est relativement simple à mettre en place et ne nécessite pas d'outils bioinformatiques poussés. En revanche, pour que la quantification soit suffisamment précise, il est nécessaire qu'un nombre de spectres MS/MS assez important soit obtenu pour chaque protéine. Pour cela, un spectromètre de masse suffisamment rapide doit être utilisé. De plus, cette méthode n'est valide que pour une gamme de concentration limitée (Bildl, Haupt et al. 2012). Par exemple, pour une protéine très peu abondante pour laquelle moins de cinq événements MS/MS ont été enregistrés, la quantification se révèle être très peu précise (Collier, Randall et al. 2011).

Ces deux méthodes de quantification sans marquage nécessitent de réaliser l'analyse de plusieurs réplicats pour une même condition. En effet, les échantillons sont traités en parallèle depuis l'extraction des protéines, la reproductibilité à chacune des étapes de préparation de l'échantillon doit être correcte, et il est nécessaire de réaliser plus de réplicats.

Contrairement aux approches avec marquage, les approches de quantification sans marquage sont d'un moindre coût.

4. L'analyse bio-informatique des données de protéomique quantitative sans marquage par extraction du signal MS

Les approches d'extraction du signal MS nécessitent l'utilisation de logiciels permettant d'extraire les intensités des signaux peptidiques et de les relier aux identifications correspondantes. Pour cela, de nombreux logiciels bio-informatiques ont été développés, comme Maxquant (Cox and Mann 2008) et MFPaQ (Bouyssie, Gonzalez de Peredo et al. 2007), développé au laboratoire.

L'extraction des signaux MS nécessite un alignement au préalable des temps de rétention de tous les peptides entre les runs des conditions à comparer, puis d'associer ces signaux MS aux identifications des peptides (données par le logiciel Mascot par exemple).

Deux approches sont possibles pour comparer les signaux MS de plusieurs conditions en « label-free ».

La première consiste à utiliser comme point de départ les résultats d'identification MS/MS. A partir de ces résultats, les signaux MS des peptides sont extraits sous forme d'un appel d'ion ou XIC (« Extracted ion chromatogram »). En effet, l'intensité du signal MS d'un peptide qui émerge de la colonne chromatographique peut être suivie au cours du temps. L'aire sous la courbe du pic chromatographique est le courant ionique extrait (XIC) et est proportionnelle à l'abondance du peptide dans l'échantillon (Ong and Mann 2005). Cette méthode consiste donc à extraire l'intensité du signal pour chaque peptide identifié, de m/z connu.

La seconde approche est centrée sur l'établissement de cartes LC-MS. Les logiciels vont définir des empreintes MS des peptides sur la base de la modélisation théorique des massifs isotopiques peptidiques et des profils d'élution de ces peptides. La quantification relative est

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

réalisée en alignant les cartes et en comparant les aires de ces empreintes peptidiques dans les différents échantillons. Ce n'est que dans un second temps que le lien avec les interprétations peptidiques est réalisé.

III. Les méthodes de quantification absolue

Les méthodes de quantification absolue permettent de déterminer la quantité absolue d'une ou de plusieurs protéines dans un échantillon d'intérêt.

Pour obtenir une quantification absolue des protéines, il est possible d'ajouter dans l'échantillon des peptides marqués (dont on connaît précisément la quantité ajoutée) qui correspondent aux peptides endogènes à quantifier qui sont eux, non marqués. La quantification se fait alors en comparant les signaux des peptides non marqués et des peptides marqués (Gerber, Rush et al. 2003, Brun, Dupuis et al. 2007).

B. Résultats

I. Conditions analytiques

Nous avons tout d'abord essayé d'évaluer une éventuelle expression différentielle des protéines entre la souche sauvage, la souche mutante $\Delta Rv1002c$ et la souche complémentée de *Mycobacterium tuberculosis*, à partir de culture réalisées dans un milieu de culture 7H9 ADC. Pour ces trois souches, nous avons comparé l'expression de ces protéines à deux temps de culture : un temps de culture correspondant à une DO de 0,5 et un temps de culture de 6 semaines. De plus, nous avons essayé d'évaluer les expressions des protéines sécrétées dans le milieu de culture mais aussi des protéines issues de la lyse des bactéries.

Etant donné que le nombre total de réplicats à analyser dans cette étude est relativement important (figure 45), nous avons choisi de ne pas réaliser de fractionnement des protéines sur gel pour chacun des échantillons. En effet, pour n'avoir à analyser qu'un seul échantillon de digestat peptidique par réplicat, nous avons déposé chacun des échantillons sur un gel SDS-Page, nous avons laissé migrer les protéines jusqu'à 1 cm après la frontière entre les gels de concentration et de séparation. Nous avons ensuite stoppé la migration, découpé une seule bande dans le gel avant de réaliser le traitement biochimique classique. Ainsi, avec l'obtention d'un digestat peptidique par réplicat, nous avons 36 échantillons à analyser pour cette étude (figure 45).

Nous aurions pu réaliser un fractionnement des protéines afin d'obtenir plusieurs digestats peptidiques par réplicat. Ceci nous aurait permis d'obtenir une couverture analytique plus

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

importante, avec un nombre de protéines identifiées et quantifiées plus important. Cependant le nombre d'échantillons à analyser aurait été trop important.

Afin d'essayer d'obtenir une quantification d'un nombre important de protéines en une seule analyse, nous avons décidé d'analyser chacun des digestats peptidiques sur un gradient chromatographique de 8h. De plus, ne pas réaliser de fractionnement des protéines en amont d'une analyse quantitative permet d'obtenir une meilleure reproductibilité des résultats entre les réplicats (Gautier, Mouton-Barbosa et al. 2012).

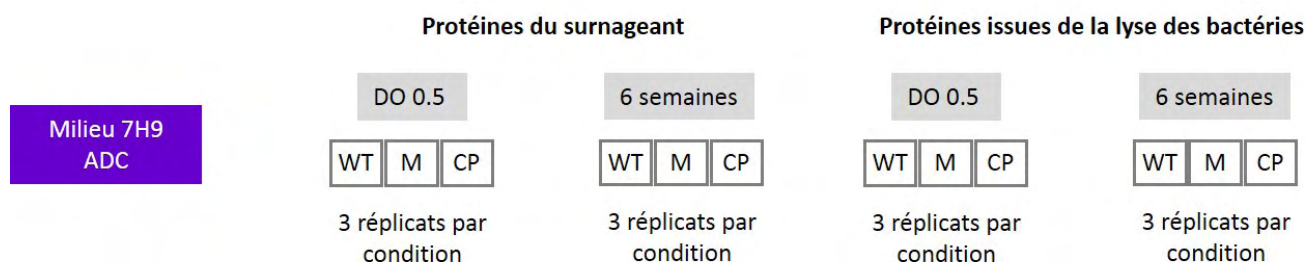


Figure 45: Conditions comparées et analysées au cours de l'étude de protéomique quantitative.

Ensuite, l'extraction des signaux MS a été réalisée grâce au logiciel MfPaQ. MfPaQ apparie tous les peptides identifiés dans les différentes conditions à comparer, puis les utilise pour aligner tous les runs en temps de rétention. Les peptides sont ensuite assignés aux protéines. Pour chaque protéine, une valeur de PAI (« Protein Abundance Index ») est calculée, qui correspond à la moyenne des intensités des 3 peptides les plus intenses de la protéine (définis sur l'ensemble des conditions comparées).

II. Protéines validées

A la suite de l'analyse nanoLC-MS/MS des digestats peptidiques et interrogation des banques de données par Mascot, une étape de validation des protéines a été faite, avec un FDR de 5%.

Dans les surnageants de culture (les deux temps de culture confondus), plus de 900 protéines ont été identifiées et validées (figure 46). Le nombre de ces protéines double entre le temps de culture de 6 semaines et celui correspondant à une DO de 0.5, ce qui semble en faveur d'une lyse des bactéries à des temps de culture important, avec libération des protéines intra cellulaires du bacille dans le surnageants de culture. En effet, les protéines sécrétées du bacille n'ont été récemment analysées et identifiées qu'au nombre de 422, des protéines intracellulaires sont alors identifiées parmi les 929 (Bell, Smith et al. 2012).

En ce qui concerne les extraits issus de la lyse des protéines, ce sont plus de 2500 protéines qui ont été validées. Or, parmi les 4012 protéines annotées de *Mycobacterium tuberculosis*,

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

ce sont 3074 protéines qu'il est possible d'observer par spectrométrie de masse (Schubert, Mouritsen et al. 2013). Au cours de cette étude, nous sommes donc parvenus à identifier plus de 80% des protéines observables par spectrométrie de masse. Notre stratégie d'analyse des protéines sans fractionnement sur gel SDS-Page nous a donc permis d'obtenir une bonne couverture analytique.

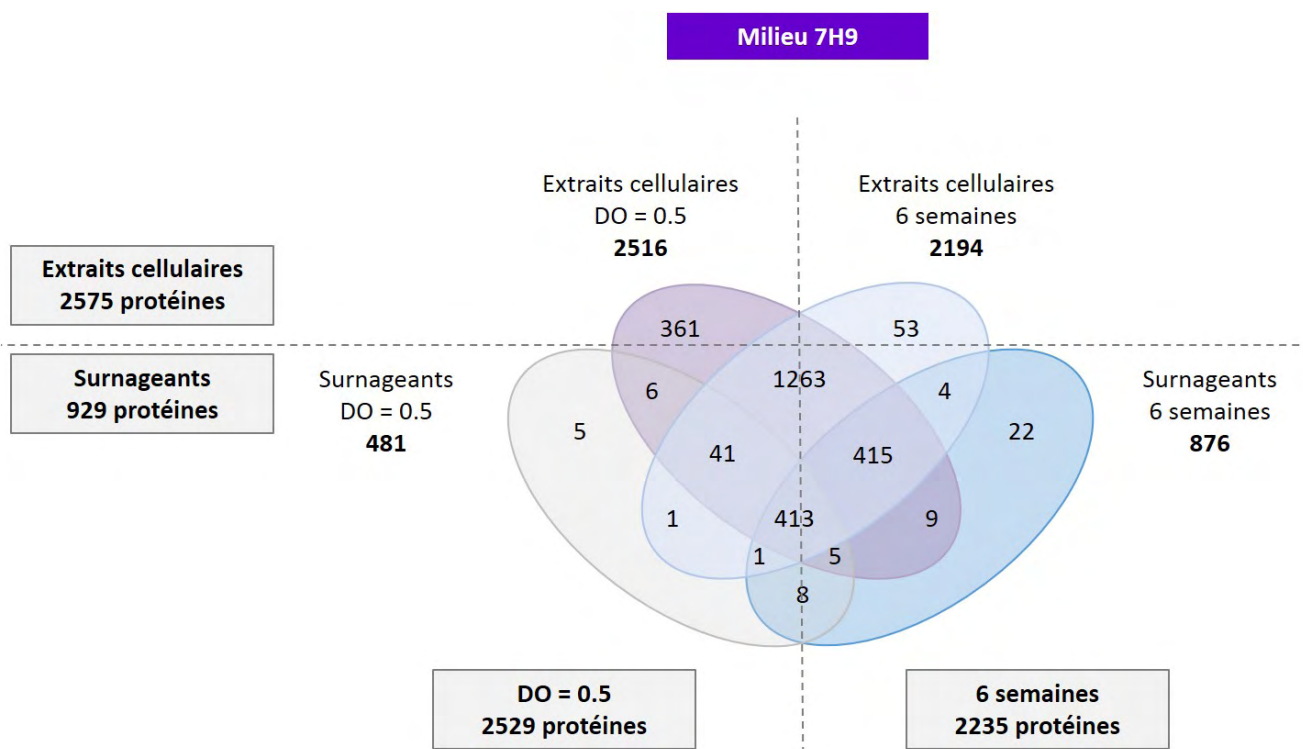


Figure 46: Nombre de protéines validées selon les conditions étudiées

III. Protéines différentiellement exprimées

Tout d'abord, à partir des valeurs des PAI obtenus pour chacune des protéines quantifiées, nous avons comparé les expressions des protéines chez la souche sauvage, la souche mutante et la souche complémentée, pour chacune des conditions étudiées (protéines du surnageant de culture ou protéines provenant de la lyse des bactéries, deux temps de culture différents).

Pour comparer l'expression des protéines dans 2 conditions différentes, les ratios des PAI de la condition 1 sur les PAI de la condition 2 à comparer ont été calculés. De plus, pour obtenir une évaluation de la significativité du résultat, les « p-value » du test de Student ont également été calculées. Ainsi, des protéines dont le ratio PAI condition 1/ PAI condition 2 est supérieur à 2 ou inférieur à $\frac{1}{2}$, avec une p-value inférieure à 0,05 sont considérées

comme variantes entre ces 2 conditions (représentées en rouge sur les graphes), dans le cas des protéines des surnageants de culture. Dans le cas des protéines issues de la lyse des bactéries, une protéine est considérée comme variante entre une condition 1 et une condition 2 si le ratio PAI condition 1/ PAI condition 2 est supérieur à 4 ou inférieur à 1/4, avec une p-value inférieure à 0.05. En effet, pour ces protéines, en considérant des protéines variantes avec pour conditions un ratio PAI condition 1/ PAI condition 2 supérieur à 2 ou inférieur à 1/2, et avec une p-value inférieure à 0.05, un grand nombre de protéines sont considérées comme variantes. Pour s'intéresser aux protéines les plus variantes, nous avons alors décidé de modifier ces valeurs.

Les graphes avec les logarithmes des ratios échantillon 1/ échantillon 2 en abscisse et logarithmes des p-value sont tracés (figures 53 et 54).

1. Protéines du surnageant de culture

a) Temps de culture correspondant à une DO de 0.5

En ce qui concerne les protéines sécrétées, pour un temps de culture correspondant à une DO de 0.5 (environ 2 semaines), très peu de protéines sont différentiellement exprimées entre la souche sauvage et la souche complémentée.

Ce sont plus de 40 protéines qui sont différentiellement exprimées entre les souches sauvage et mutante, sur 426 protéines quantifiées (figures 47 et 53).

Parmi les protéines sur exprimées chez la souche mutante, les 3 antigènes du complexe 85 (85A (fbpA, Rv3804c), 85B (fbpB, Rv1886c) et 85C (fbpC, Rv0129c)) sont retrouvés. Ces trois antigènes constituent les 3 protéines les plus abondantes au sein du surnageant de culture de *Mycobacterium tuberculosis* (De Souza and Wiker, 2011). Ces protéines sont fortement immunogéniques au cours d'infections mycobactériennes, en terme de production d'anticorps et induction de stimulation des lymphocytes T (Wiker and Harboe, 1992). De plus, 3 autres antigènes sont également sur exprimées chez la souche mutante : les antigènes mpt53 (Rv2878c), mpt64 (Rv3036c) et pstS1 ou Ag78 (Rv0934).

De plus, les protéases homologues clpP1 (Rv2461c) et clpP2 (Rv2460c), requises pour la viabilité et la croissance de la bactérie *in vitro*, ainsi que pour la croissance de la bactérie durant l'infection chez la souris sont également sur exprimées chez la souche mutante (Raju, Unnikrishnan et al. 2012).

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

La protéine DevR du régulon DosR est retrouvée sous exprimée chez la souche mutante (figure 48). Cette protéine, tout comme la majorité des protéines du régulon DosR est surexprimée par la bactérie lorsque celle-ci est en condition d'hypoxie (Chauhan, Sharma et al. 2011). De même, 4 autres protéines généralement exprimées en réponse d'une hypoxie sont surexprimées chez la souche sauvage : le facteur d'élongation Tu, la protéine HspX, une protéine universelle du stress, ainsi qu'une protéine putative.

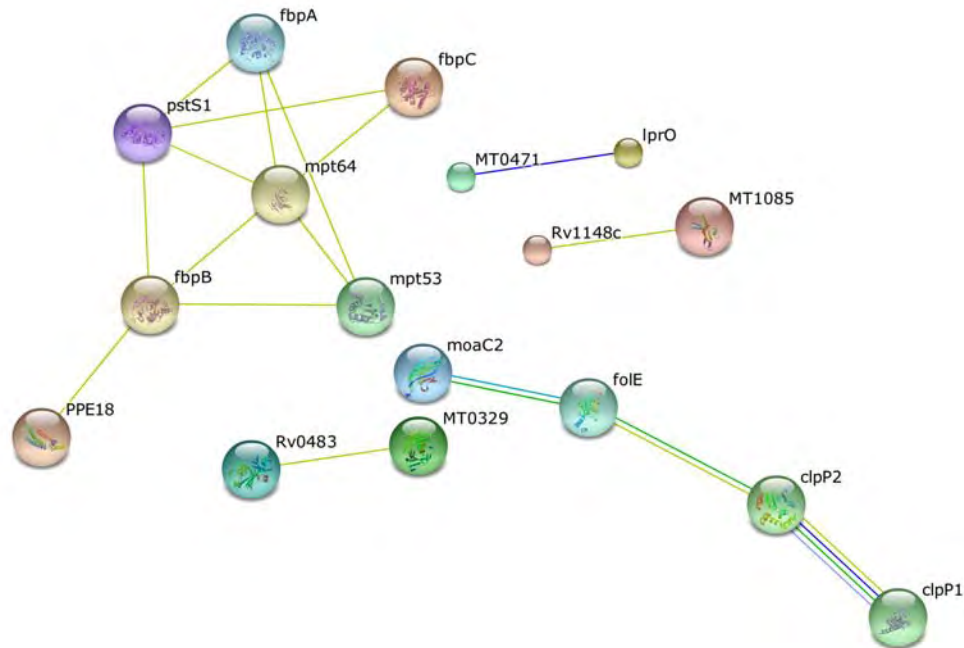


Figure 47: Réseau des protéines du surnageant de culture surexprimées chez la souche mutante par rapport à la souche sauvage, temps de culture correspondant à une DO de 0.5. Réseau obtenu avec le logiciel STRING.

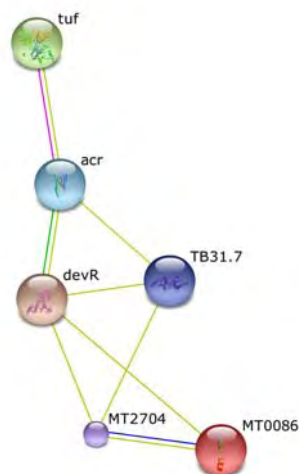


Figure 48: Réseau des protéines du surnageant de culture sous exprimées chez la souche mutante par rapport à la souche sauvage, temps de culture correspondant à une DO de 0.5. Réseau obtenu avec le logiciel STRING.

b) Temps de culture de 6 semaines

Pour les données correspondant aux protéines du surnageant de culture, après un temps de culture de 6 semaines, des problèmes d'alignement des spectres ont été rencontrés. En effet, certainement en raison d'un problème chromatographique inhérent à l'utilisation de gradients relativement longs, de gros écarts entre les temps de rétention des mêmes peptides dans les différents runs à comparer ont été observés. En raison de ces problèmes d'alignement, seulement 282 protéines ont été quantifiées (pour les 3 souches confondues) sur les 929 protéines identifiées dans les mêmes échantillons. En effet, pour qu'un peptide soit quantifié, il faut que le profil chromatographique de celui-ci soit retrouvé dans les runs chromatographiques des différentes conditions à comparer. Or, dans le cas où les temps de rétention de ce peptide sont trop différents entre les différents runs, l'outil d'extraction des signaux MS ne parvient pas trouver l'information à extraire et la quantification de ce peptide et donc de la protéine correspondante (si le même problème est retrouvé pour tous les peptides de la protéine) n'est pas possible.

Une protéine putative et la protéine pks4 (Rv1181) sont sous exprimées chez la souche mutante par rapport à la souche sauvage. La protéine ESAT6 (Rv3675), antigène stimulant les lymphocytes T, est sur exprimée chez la souche mutante par rapport à la souche sauvage (figure 54).

2. Protéines issues de la lyse des bactéries

a) Temps de culture correspondant à une DO de 0.5

Ce sont 2332 protéines issues de la lyse des bactéries, après un temps de culture correspondant à une DO de 0,5 qui ont été quantifiées (figure 53).

A un temps de culture équivalent à 2 semaines, les classes dont les protéines sont les plus différenciellement exprimées entre la souche sauvage et la souche mutante sont la classe des protéines de type transporteurs ABC, la classe des protéines de l'export et de la sécrétion bactérienne, et la classe des protéines de la phosphorylation oxydative (figures 49 et 50).

3 des 5 protéines du système Sec d'export des protéines sont sur exprimées chez la souche mutante : SecD (Rv2587c), SecE (Rv 0638) et SecY (Rv0732).

Les deux protéines du système de régulation à deux composants mprA/mprB (Rv0981 et Rv0982) sont également sur exprimées chez la souche mutante. Ce système est requis par la bactérie pour se maintenir chez l'hôte au cours de l'infection.

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

De plus, 3 protéines du complexe mce1 sont également sur exprimées chez la souche mutante : mce1C (Rv0171), mce1D (Rv0172) et mce1F (Rv0174). Ce complexe a un rôle dans l'activation des gènes impliqués dans le trafic des substrats pendant la phase initiale d'infection.

Deux facteurs de virulence mpt51 (Rv3803c) et mpt64 (Rv1980c) sont sous exprimées chez la souche mutante (figure 51).

Term	p-value	n°
ABC transporters	1.330e-4	16
Protein export	3.309e-2	6
Oxidative phosphorylation	5.600e-2	10
Bacterial secretion system	6.370e-2	5
Biotin metabolism	1.360e-1	4
Styrene degradation	8.539e-1	2

Figure 49: Les principales classes de protéines issues de la lyse des bactéries qui sont différemment exprimées entre la souche sauvage et la souche mutante, à un temps de culture correspondant à une DO de 0,5

b) Temps de culture correspondant à un temps de culture de 6 semaines

Ce sont 2332 protéines issues de la lyse des bactéries, après un temps de culture de 6 semaines, qui sont quantifiées.

De la même façon que pour un temps de culture correspondant à une DO de 0.5, les facteurs de virulence mpt51 (Rv3803c) et mpt64 (Rv1980c) sont sous exprimées chez la souche mutante, ainsi que les antigènes mpt53 (Rv2878c), mpt63 (Rv1926c), l'antigène 85A (Rv3804c) et la protéine glycosylée Apa (Rv1860) (figure 52).

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

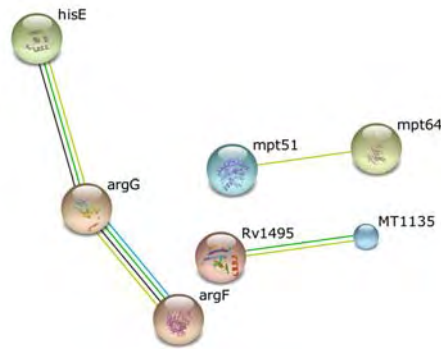


Figure 51: Réseau des protéines issues de la lyse des bactéries sous exprimées chez la souche mutante par rapport à la souche sauvage, temps de culture correspondant à une DO de 0.5. Réseau obtenu avec le logiciel STRING.

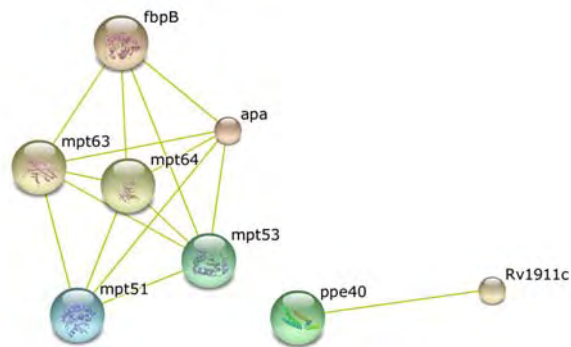


Figure 52: Réseau des protéines issues de la lyse des bactéries sous exprimées chez la souche mutante par rapport à la souche sauvage, temps de culture de 6 semaines. Réseau obtenu avec le logiciel STRING.

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

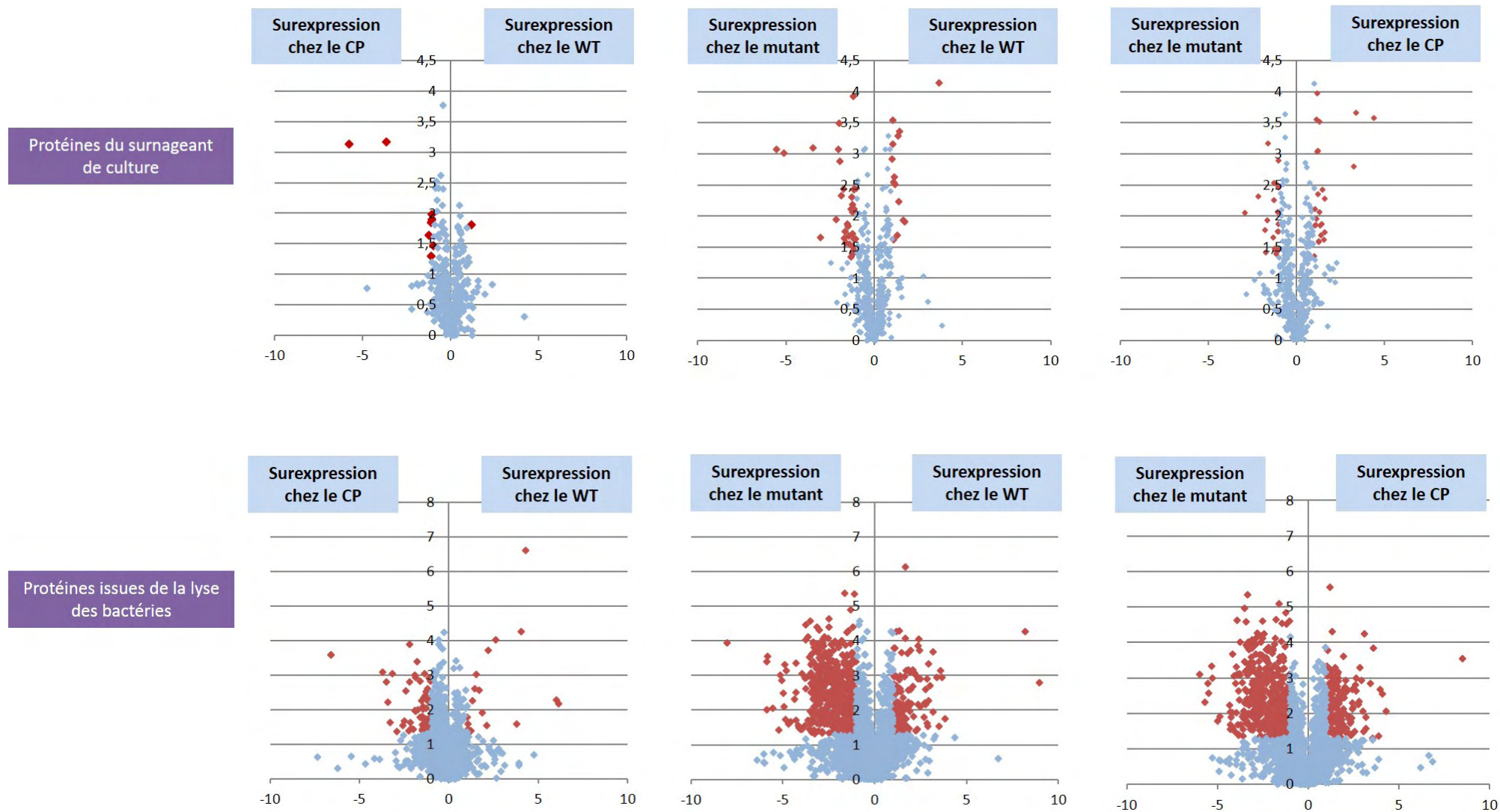


Figure 51: Représentation graphique de l'analyse quantitative des protéines sécrétées et issues de la lyse des bactéries de la souche sauvage, de la souche mutante et de la souche complémentée de *Mycobacterium tuberculosis*, à un temps de culture correspondant à une DO de 0,5.

RESULTATS deuxième partie: Recherche des voies métaboliques de Mtb impactées par un arrêt du processus de glycosylation

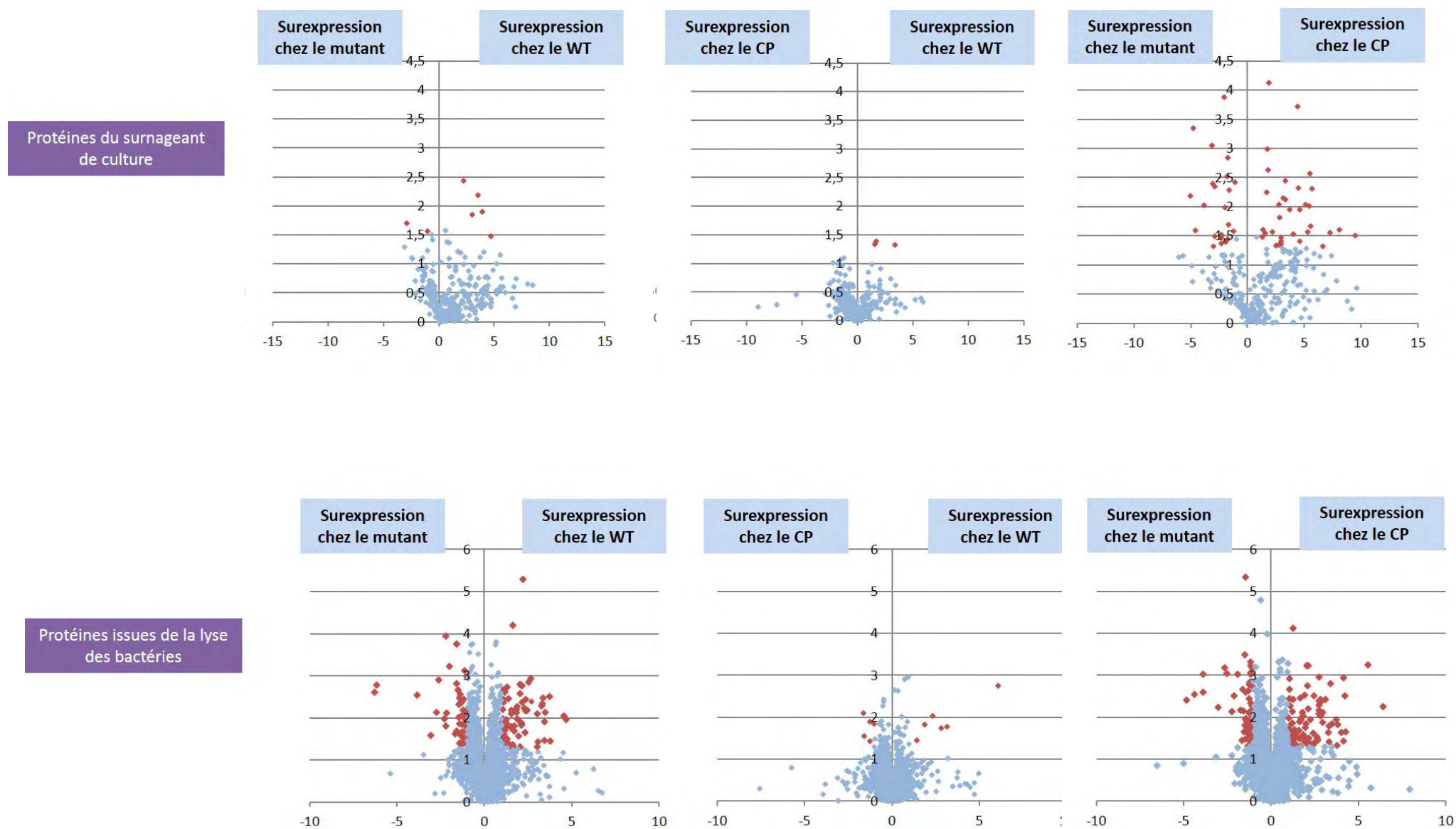


Figure 54: Représentation graphique de l'analyse quantitative des protéines sécrétées et issues de la lyse des bactéries de la souche sauvage, de la souche mutante et de la souche complémentée de *Mycobacterium tuberculosis*, à un temps de culture de 6 semaines

IV. Conclusion

Ces résultats sont des résultats préliminaires. En effet leur analyse est encore en cours. Cependant, ils font apparaitre des différentiels d'expressions de certaines protéines entre les souches sauvage et mutante qui pourraient être spécifiquement associés à l'invalidation du gène Rv1002c de la protéine-*O*-Mannosyl-transferase.

En parallèle de cette étude, une analyse transcriptomique dans les mêmes conditions a également été réalisée. Ces données sont également encore en cours d'analyse et seront confrontées aux données de protéomique quantitative afin de pouvoir mettre en évidence des altérations métaboliques ou physiologiques de la bactérie pouvant expliquer le phénotype de virulence atténué observé pour la souche mutante.

RESULTATS troisième partie: Mise en place de méthodes d'analyse par spectrométrie de masse pour la caractérisation de la glycosylation de la Fasciline de *Mycobacterium smegmatis*

A. Contexte

La glycosylation de la Fasciline a été découverte au laboratoire, avant ce travail de thèse. En effet, suite à la caractérisation des phénotypes de la souche sauvage de *Mycobacterium smegmatis* et de la souche mutante délétée du gène Δ MsMeg 5447, une analyse protéomique différentielle des protéines des surnageants de ces deux souches avait été réalisée. C'est à partir de cette étude que la liste des glycoprotéines de *Mycobacterium smegmatis* avait été établie (voir résultats partie A, paragraphe I.D.2).

En effet, sur le gel SDS-Page sur lequel ont été déposées les protéines sécrétées des souches sauvage, mutante et complétée (figure 55A), une bande intense apparaît dans les 3 pistes et correspond à une protéine d'un poids moléculaire apparent entre 25 et 30 kDa, avec une diminution de celui-ci dans le cas de la souche mutante.

Après excision de cette bande, digestion trypsique, analyse nanoLC-MS/MS des peptides, et interrogation des banques de données, la protéine a été identifiée comme étant la Fasciline. Pour pouvoir mieux caractériser cette protéine, la protéine recombinante avec un tag histidine a été produite. Après purification, une hydrolyse de la protéine a été réalisée afin de libérer les motifs saccharidiques et en étudier le contenu par électrophorèse

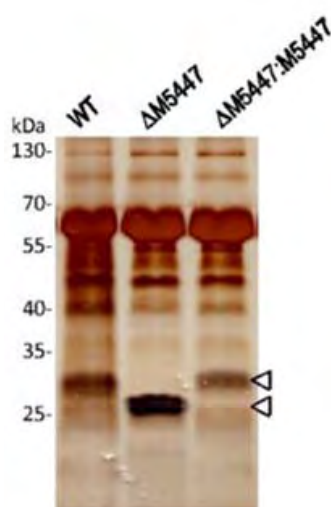


Figure 55A: Gel SDS-Page réalisé avec les protéines du surnageant de culture de la souche sauvage, de la souche mutante et de la souche complétée de *Mycobacterium smegmatis*

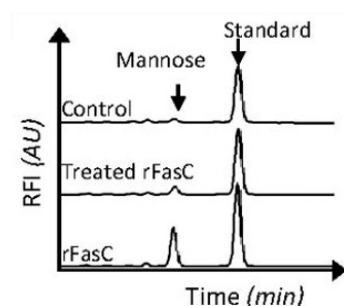


Figure 55B: Analyse par électrophorèse capillaire des monosaccharides libérés à l'issue d'une hydrolyse. Treated rFasC = Fasciline traitée à l' α -mannosidase.

capillaire (figure 55B). La présence de mannoses sur la protéine a été confirmée.

Afin de mettre en place des stratégies d'analyse par spectrométrie de masse de glycoprotéines mycobactériennes, en vue d'une analyse approfondie des glycoprotéines de *Mycobacterium tuberculosis* nouvellement identifiées, il a été décidé d'utiliser cette protéine comme protéine modèle pour une étude approfondie de la glycosylation et en particulier des glycosites. En effet, cette protéine, avec un tag histidine pour sa purification sur colonne de nickel, possède 24 thréonines et 9 sérines (figure 56 en bleu et violet), dont 21 sérines et thréonines qui possèdent de bonnes probabilités de glycosylation selon le logiciel NetOGlyc (en bleu). Cette protéine semble donc être un bon modèle pour l'étude de la glycosylation des protéines mycobactériennes. A partir d'une culture d'une semaine d'une souche de *Mycobacterium smegmatis* produisant cette protéine avec le tag histidine, nous avons récupéré le surnageant de culture et purifié la protéine sur colonne de Nickel.

MHTRKTLGAAAAIAAIATSLPLAVTAYAEPE**TTTE**AEPTVEIPDPQGPGCDDFKKA
VP**SW**KALNDVPV**SQ**ALAAIPEI**ST**FNAAV**SG**GLNPAVNIDPVLNDNGPYVVFAPT**N**
EAFAKLAPGQLDAIAADPAAL**TS**LDYYHAFLGLLGNDDVKGQRP**TQ**QGAIEIKV**TG**
KGGDIEVND**TA**KLVCGAIHAQNARIYLID**TV**LDPNNPPEAL**TP**TV**SGTSTTPTTTT**
TTEPAPAEAPE**SAET**PEAMPAADAPI**GS**HHHHHH

Figure 56: Séquence de la Fascicline recombinante. En gras et en bleu: sérines et thréonines ayant de bonnes probabilités de glycosylation selon NetOGlyc. En gras et en violet: sérines et thréonines n'ayant pas de bonnes probabilités de glycosylation selon NetOGlyc. En souligné: le peptide C-terminal trypsique.

Nous allons donc évaluer l'apport de différentes techniques d'analyse par spectrométrie de masse pour une caractérisation complète de la glycosylation de cette protéine. La première analyse que nous avons réalisée est l'analyse de la protéine entière afin de déterminer le nombre d'hexoses portés par la protéine, ainsi que d'autres PTMs éventuellement présentes.

B. Détermination des modifications post-traductionnelles de la protéine

I. Mise en place d'une analyse de la protéine entière pour la détermination des protéoformes

Afin d'obtenir le profil de la Fascicline recombinante par spectrométrie de masse (visualisation des différentes protéoformes), nous avons travaillé en couplage chromatographie liquide – spectrométrie de masse sur l'Orbitrap Velos.

Parmi les différents spectres MS enregistrés au cours de l'analyse, ceux dans lesquels apparaissent des pics correspondant à des valeurs de m/z théoriques de la protéine ont été moyennés (figure 57). Un spectre complexe est obtenu, avec une distribution de charge de 12 à 21. Des pics très intenses apparaissent correspondants à différents états de charge, ainsi que des pics beaucoup moins intenses.

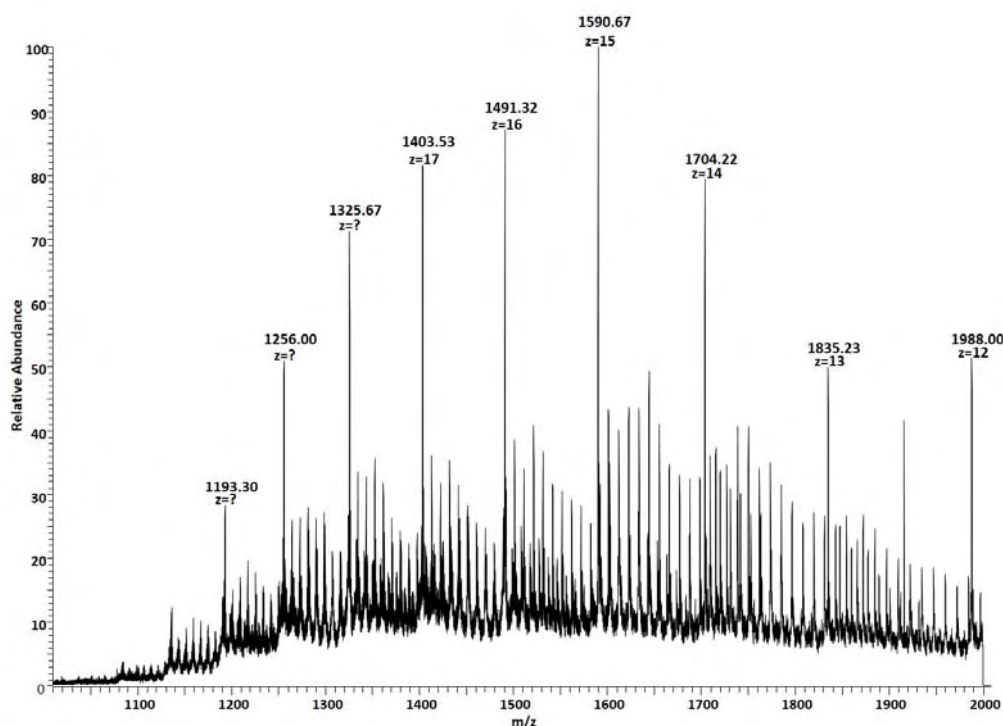


Figure 57: Spectre MS moyen de la fascicline recombinante

Après déconvolution de ce spectre moyenné (obtention des valeurs de masse à partir des valeurs de masse/charge), un pic intense correspond à la protéine non modifiée et des pics moins intenses correspondent aux différents états de glycosylation de la protéine (figure 58). Un nombre d'hexose maximal de 17 est détecté. Un tel profil de glycosylation n'a jamais été mis en évidence auparavant pour une protéine mycobactérienne.

De plus, une autre modification est détectée. Il s'agit de l'oxydation de la méthionine du peptide C-terminal. Cette oxydation peut être une modification post-traductionnelle ou peut être une oxydation chimique introduite au cours du traitement de l'échantillon, la seconde hypothèse étant plus vraisemblable.

L'analyse de la protéine entière nous a donc permis de confirmer sa glycosylation et de déterminer le nombre d'unités hexoses fixées sur celle-ci. Pour essayer d'identifier les

RESULTATS troisième partie : Mise en place de méthodes d'analyse MS pour la caractérisation de la Fascicline

peptides glycosylés de la protéine, nous avons ensuite commencé par réaliser une analyse nanoLC-MS/MS des peptides obtenus après digestion à la trypsine.

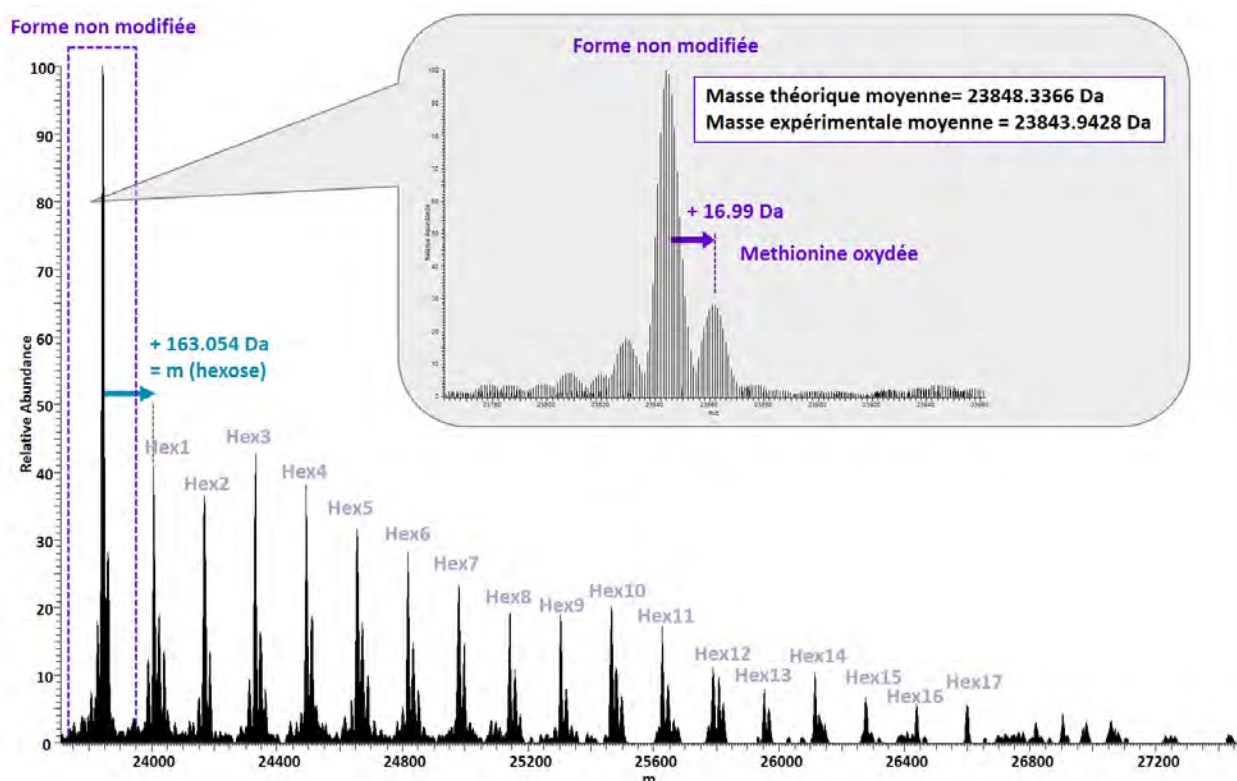


Figure 58 : Spectre MS déconvolué de la fascicline recombinante. Un nombre d'hexose maximal de 17 est détecté.

Un écart de près de 5 Da est obtenu (soit 0.02%) entre les masses moyennes théoriques et expérimentales. De plus, une masse de 163 Da au lieu de 162 Da est obtenue pour le premier motif hexose, et une masse de près de 17 Da au lieu de 16 Da est obtenue pour l'oxydation de la méthionine. Les masses moyennes qui apparaissent sur ce spectre sont en effet issues de la déconvolution du spectre moyenné de la protéine et cette opération peut entraîner des imprécisions de mesure.

II. Recherche des peptides modifiés via une approche de type « bottom-up »

Avant ce travail de thèse, une analyse préliminaire par nanoLC-MS/MS des peptides tryptiques de la protéine non recombinante avait été réalisée. Au cours de cette étude, deux peptides N-terminaux 1-26 et 1-27 (qui diffèrent par la présence d'une lysine côté C-terminal non clivée) avaient été identifiés comme peptides glycosylés, en recherchant les spectres MS/MS de peptides possédant une signature de peptide glycosylé. Le peptide 1-26 avait été identifié avec 4 hexoses, contre 3 hexoses pour le peptide 1-27.

De la même façon que précédemment, après analyse nanoLC-MS/MS des peptides tryptiques de la protéine recombinante, ceux possédant une signature de peptide glycosylé ont été retenus.

Les peptides N-terminaux déjà identifiés précédemment à partir de la protéine non recombinante ont été ici retrouvés, avec de 1 à 3 hexoses sur le peptide 1-26, et de 1 à 4 hexoses sur le peptide 1-27 (figure 59). Ce résultat diffère de celui obtenu au cours de l'analyse de la protéine non recombinante. En effet, la distribution des hexoses sur la protéine non recombinante et sur la protéine recombinante peut être différente.

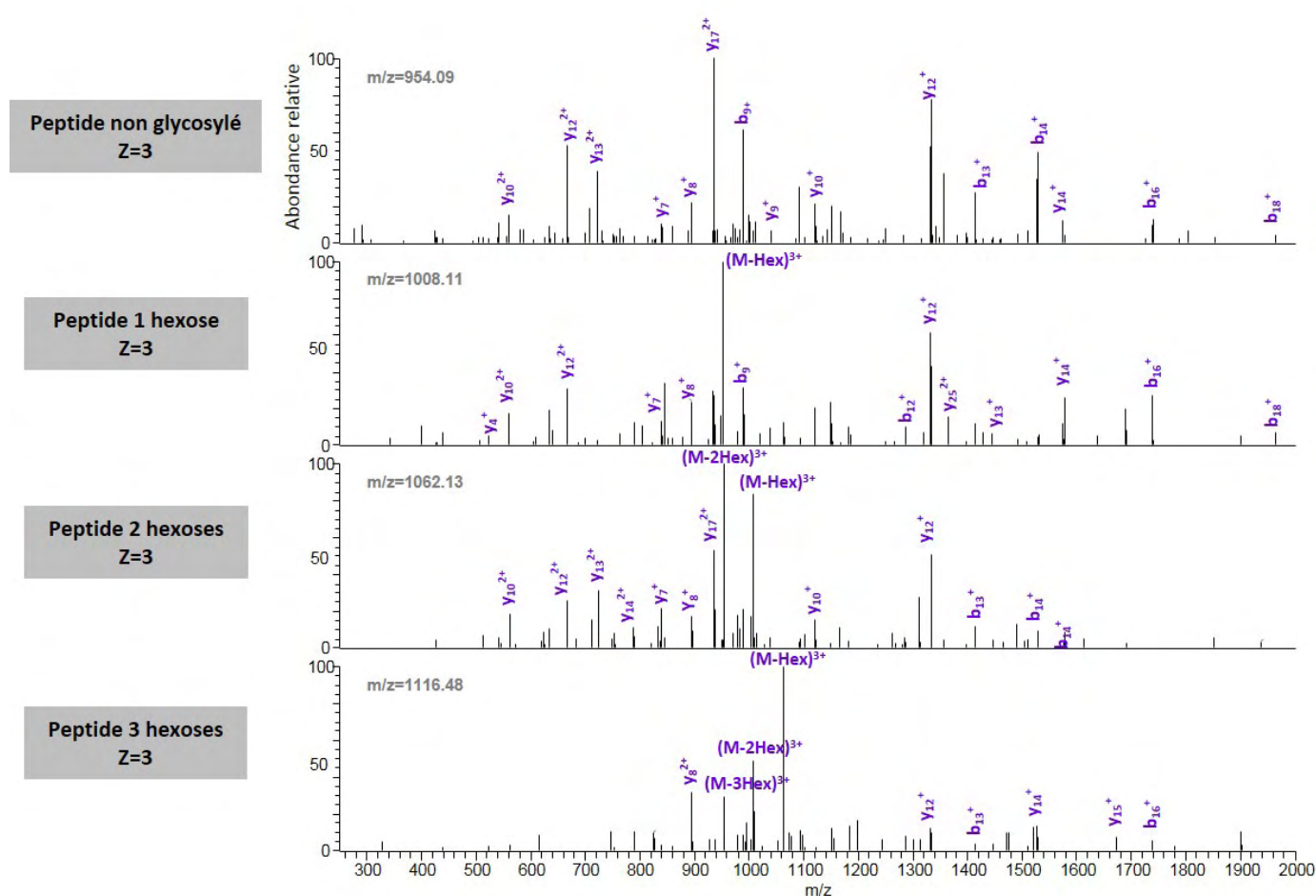


Figure 59: Spectres MS/MS CID des formes non glycosylée, mono-, di- et tri- glycosylées du peptide N-terminal 1-26 de la fascicline recombinante. Au sein des spectres des formes mono-, di- et tri- glycosylées, des ions intenses au sein des spectres correspondent aux ions provenant des pertes des différents hexoses à partir des ions parents.

En revanche, aucun spectre MS/MS n'a été obtenu pour le peptide C-terminal tryptique. En effet, la couverture de la séquence de la protéine n'est pas de 100% mais de 46%, deux peptide tryptiques de masses importantes dépourvus de sites de clivage internes n'ont pas été séquencés au cours de cette analyse (figure 60 et tableau 13).

EPETTTEAEPTVEIPDPQGPQCDDFKKAVPSWKALNDVPVSQALAAIPEISTFN
AAVSGGLNPAVNIDPVLNDNGPYVVFAPTNEAFKALAPGQLDAIAADPAALTSLD
YYHAFLGLLGNDVDKQGRPTQQGAEIKVTGKGGDIEVNDTAKLVCGAIHAQNA
RIYLIDTVLDPNNPPEALTPTVSGTSTTPTTTTETTEPAPAEAPESAETPEAMPAA
DAPIGTSHHHHHH

Figure 60: Couverture de séquence de la protéine. En rouge : parties de la protéine couverte par l'analyse CID des peptides.

Position	Peptide	Masse	Charge	PTM
1-26	EPETTTEAEPTVEIPDPQGPQCDDFK	3182,36	2;3	Hex1, Hex2, Hex3, Cam
1-27	EPETTTEAEPTVEIPDPQGPQCDDFKK	3310,45	2;3;4	Hex1, Hex2, Hex3, Hex4, Cam
27-33	KAVPSWK	814,47	2	
28-33	AVPSWK	686,37	2	
84-118	LAPGQLDAIAADPAALTSLDYYHAFLGLLGNDVDK	3627,85	3;4;5;6	
119-130	GQRPTQQGAEIK	1311,69	2;3;4	
119-134	GQRPTQQGAEIKVTGK	1696,92	2;3;4	
122-130	PTQQGAEIK	970,51	2	
122-134	PTQQGAEIKVTGK	1355,74	3	
131-145	VTGKGGDIEVNDTAK	1503,74	2;3	
135-145	GGDIEVNDTAK	1117,52	2;3	
135-157	GGDIEVNDTAKLVCGAIHAQNAR	2409,17	2;3;4	Cam
146-157	LVCGAIHAQNAR	1322,69	3	Cam

Tableau 13: Liste des peptides tryptiques fragmentés et identifiés au cours de l'analyse (2 clivages manqués tolérés).

Cam = carbamidométhylation des cystéines

Or, c'est sur ce peptide que sont localisées la plupart des sérines et thréonines ayant de fortes probabilités de glycosylation de la protéine. Pour essayer d'obtenir des informations sur ce peptide, nous avons alors réalisé des appels d'ions sur les différents états de charge possible de ce peptide. Nous avons ainsi pu extraire de tous les spectres MS enregistrés au cours de l'analyse les spectres MS de celui-ci. Ces spectres ont été enregistrés sur un temps d'élution relativement long et ils ont été moyennés. L'intensité des pics correspondant aux différents états de charge (de 6 à 8) et aux différentes formes de ce peptide est relativement faible, avec de nombreux autres ions présents sur le spectre, ce qui peut expliquer que ce peptide n'ait pas été sélectionné par le spectromètre de masse pour la fragmentation (figure 61).

RESULTATS troisième partie : Mise en place de méthodes d'analyse MS pour la caractérisation de la Fascicline

Après déconvolution de ce spectre moyenné, un spectre complexe est obtenu, avec 20 états de glycosylation du peptide (figure 62). Ce sont les formes avec un état de glycosylation de 3 à 13 qui sont les plus abondantes, et non la forme non glycosylée comme c'est le cas après analyse de la protéine entière. Cependant, les ions d'intérêt au sein du spectre MS moyenné qui ont subi la déconvolution ont une intensité très faible, les intensités des différentes formes après déconvolution ne sont peut-être pas représentatives des intensités réelles.

De plus, un zoom sur la région de la première forme qui correspond à la forme non modifiée permet de mettre en évidence l'oxydation de la méthionine présente dans le peptide que nous avons également déterminée au cours de l'analyse de la protéine entière, ainsi que la présence d'une autre modification d'une masse de 57 Da (figure 63). Nous n'avons pas réussi à déterminer la nature de cette dernière modification, qui n'apparaît pas lors de l'analyse de la protéine entière.

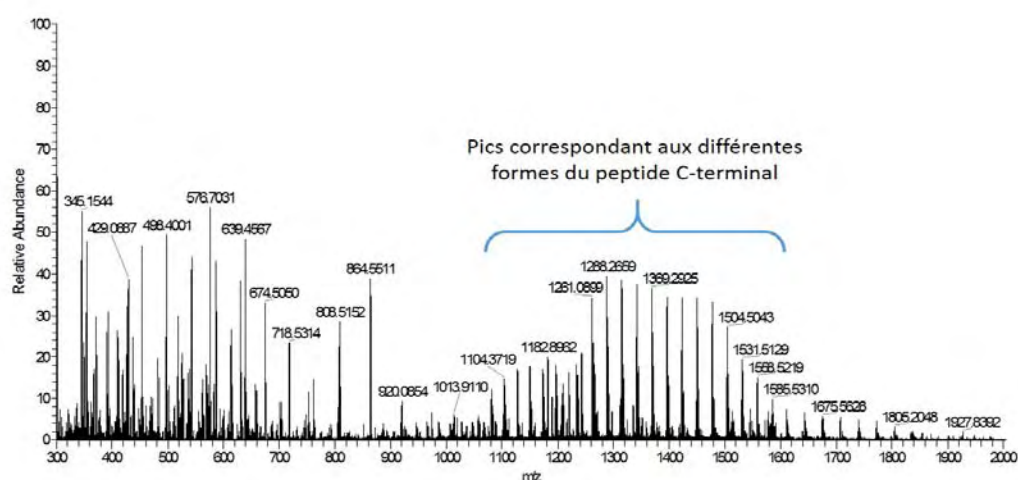


Figure 61: Spectre moyen de tous les spectres MS dans lesquels apparaissent les différents pics du peptide C-terminal.

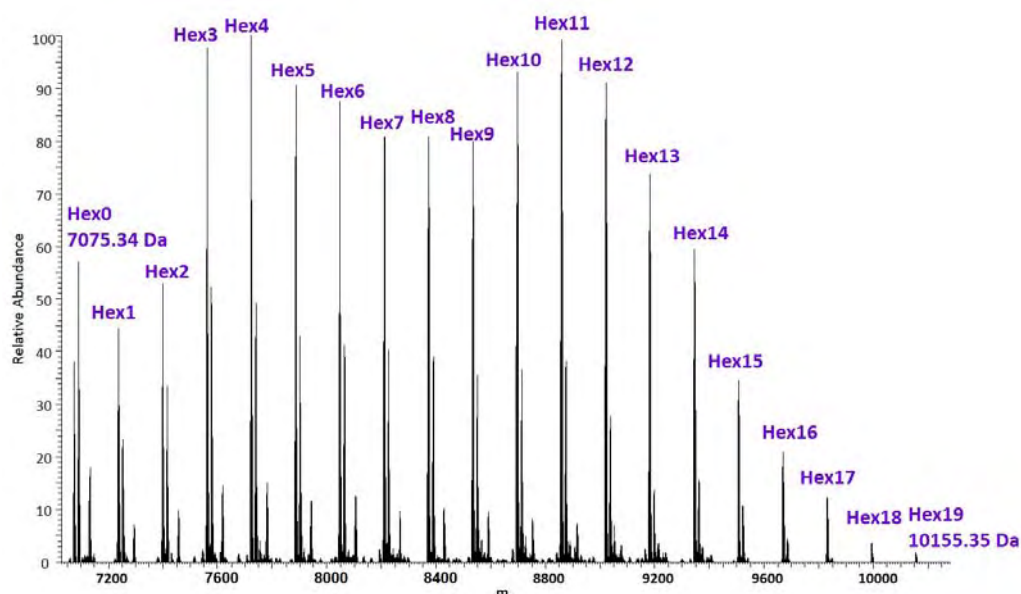


Figure 62: Spectre MS déconvolué du peptide C-terminal de la Fascicline

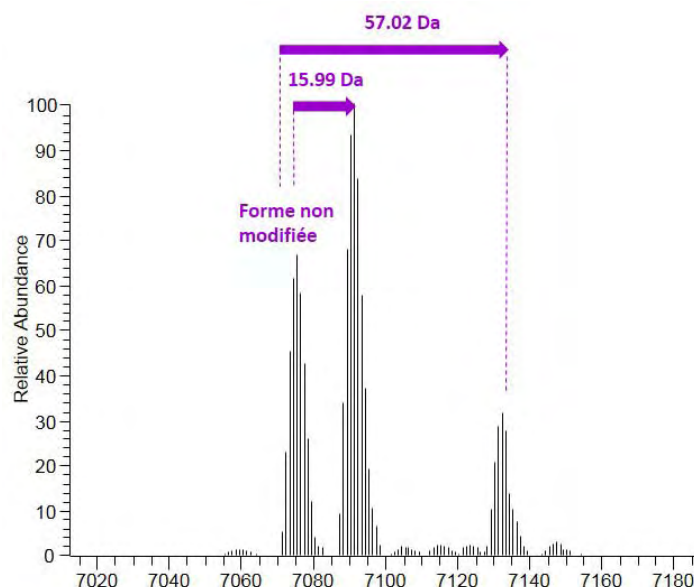


Figure 63: Zoom sur la forme non-glycosylée du peptide C-terminal. Deux modifications sont visibles: l'oxydation de la methionine (masse de 15.99 Da) et une autre modification d'une masse de 57.02 Da.

Outre les peptides N- et C-terminaux, aucun autre peptide glycosylé n'a été détecté au cours de l'analyse nanoLC-MS/MS des peptides tryptiques de la protéine.

Afin de déterminer ensuite les sites de glycosylation de la protéine, nous avons envisagé plusieurs stratégies, au niveau protéique ou au niveau peptidique.

C. Caractérisation des sites de glycosylation de la protéine

Afin de déterminer les sites de glycosylation de la protéine, qui se trouvent principalement dans la région C-terminale de la protéine, nous avons tout d'abord envisagé d'utiliser une approche complémentaire à l'approche « bottom-up », l'approche « top-down » qui consiste en la fragmentation des protéines entières. En effet, la combinaison de plusieurs modes de fragmentation de la protéine peut permettre l'identification de fragments glycosylés, ou la détermination des sites de glycosylation si les sites de fragmentation sont suffisamment nombreux (Gault, Malosse et al. 2013)

Ensuite, pour confirmer les sites de glycosylation déterminés avec l'approche « top-down », ou pour obtenir une localisation plus fine des sites de glycosylation, nous avons également choisi de tester l'approche « bottom-up » avec un autre mode de fragmentation, le mode ETD, particulièrement adapté pour la détermination des sites de glycosylation dans le cadre de l'étude de modifications post-traductionnelles labiles (voir introduction première partie , paragraphe B.II.4.).

I. Mise en place d'une analyse de type « top-down » pour une caractérisation des fragments modifiés de la protéine

1. Conditions analytiques

Afin d'essayer d'obtenir une bonne fragmentation de la protéine pour parvenir à localiser les motifs hexoses sur celle-ci, nous avons testé différents modes de fragmentation : le mode ETD adapté pour la recherche de modifications post-traductionnelles, le mode HCD, ainsi que le mode de fragmentation en source SID (voir introduction partie A, paragraphe III.B.1). En effet, l'utilité du mode ETD lors de la fragmentation de protéines pour l'étude des PTMs a bien été documenté, contrairement à l'utilité des modes SID et HCD. Pour évaluer le comportement des motifs hexoses portés par les protéines au cours d'une fragmentation avec les modes SID et HCD, nous avons décidé de les tester.

Pour les essais de fragmentation en source, plusieurs voltages de 25 à 80V ont été testés et la valeur de 80V a tout d'abord été choisie pour obtenir une fragmentation satisfaisante de la protéine. Les fragments formés en source et les protéoformes entières non fragmentées sont ensuite analysés dans l'Orbitrap avec une résolution de 100 000.

Pour les essais de fragmentation HCD et ETD, un ion précurseur correspondant à une glycoforme de la protéine et à un état de charge donné est sélectionné avec une fenêtre de sélection de 10 m/z (fenêtre de sélection centrée sur l'ion d'intérêt avec une tolérance de masse de 5 m/z du côté des masses inférieures ainsi que du côté des masses supérieures) et fragmenté avec le mode de fragmentation choisi. Le précurseur et les fragments sont analysés dans l'Orbitrap, avec une résolution de 15 000 pour le précurseur et de 100 000 pour les fragments (figure 64).

A la suite de l'acquisition des données, celles-ci sont retraitées avec le logiciel ProSight PC de Thermo.

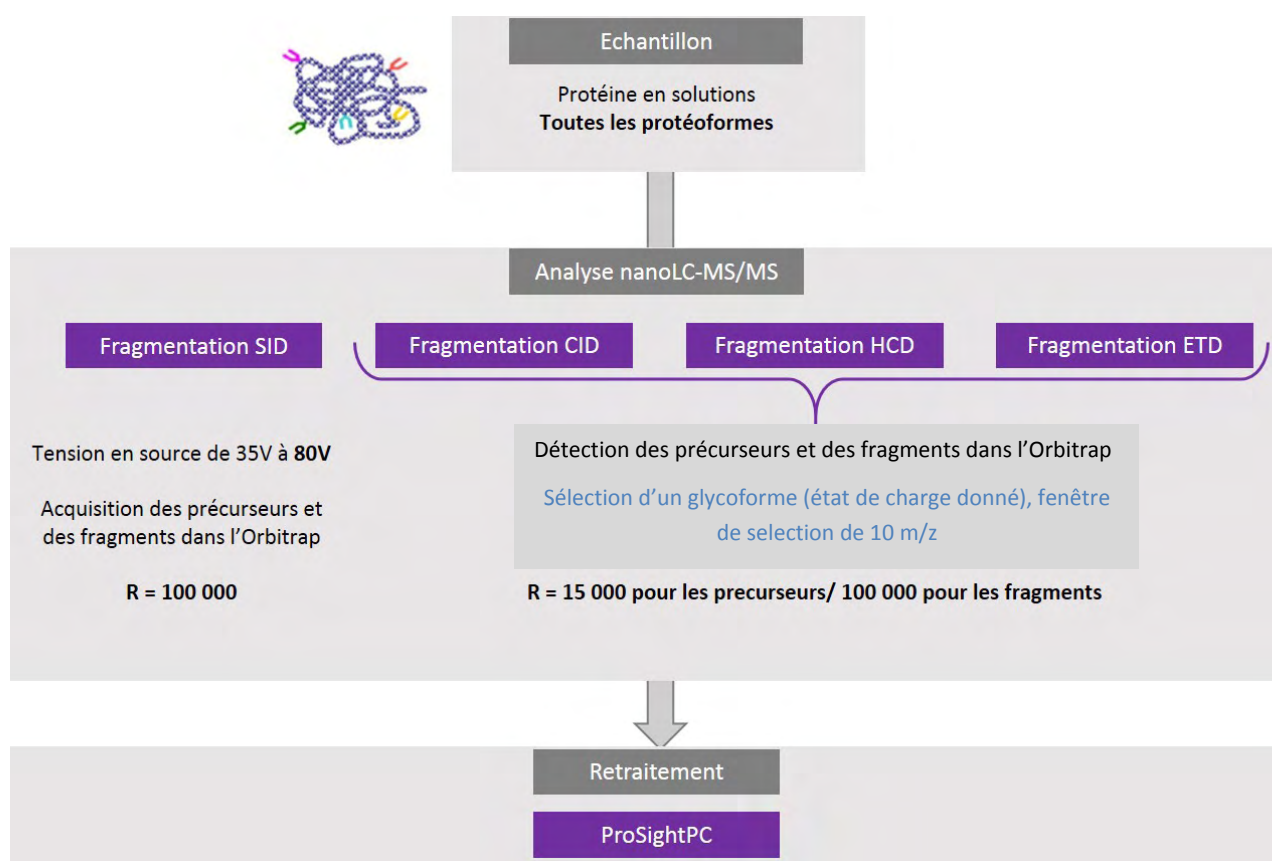


Figure 64: A partir de l'échantillon protéique, les différentes méthodes analytiques utilisées selon les modes de fragmentation et retraitement des données avec le logiciel ProSightPC

2. Résultats – fragmentation SID

Au sein du logiciel ProSight PC est implémenté le programme de déconvolution Xtract précédemment utilisé pour déconvoluer les spectres MS moyennés. A partir des spectres MS enregistrés dans l'Orbitrap contenant les massifs isotopiques des ions correspondant aux protéoformes non fragmentées ainsi qu'aux fragments, l'outil Xtract permet d'obtenir la liste des masses monoisotopiques de toutes les espèces présentes dans le spectre. ProSight PC confronte ensuite cette liste de masses monoisotopiques expérimentales avec une liste de masses monoisotopiques théoriques. Cette liste de masses monoisotopiques théoriques peut être calculée à partir d'une banque de données protéiques ou à partir de la séquence d'une protéine entrée manuellement. La protéine que nous analysons étant une protéine taguée non présente dans les banques de données, c'est la deuxième option que nous avons utilisée. Nous avons choisi une tolérance de 5ppm entre les masses expérimentales et théoriques pour cette recherche.

Prosight PC nous a tout d'abord permis de rechercher des fragments portant de 0 à 10 hexoses, avec ou sans oxydation de la méthionine dans la région C-terminale. En effet, ce sont toutes les protéoformes de la protéine qui ont été potentiellement fragmentées, puisque aucune sélection d'un ion précurseur n'est réalisée dans le spectromètre de masse. Selon le comportement des hexoses face à l'énergie de collision appliquée, les hexoses portés par certains fragments peuvent éventuellement être détectés sur certains fragments ou peuvent avoir été éliminés.

71 fragments portant de 0 à 10 hexoses ont été détectés. Comme il a été vu précédemment dans le cas de la fragmentation en source (voir introduction partie A, paragraphe III.B.1), ce sont principalement des fragments aux extrémités N- et C-terminales qui sont obtenus (figure 65). 27 fragments sans hexoses, avec méthionine oxydée (fragments en gras) ou non oxydée (fragments d'épaisseur normale) ont été détectés. Ce sont donc 44 fragments glycosylés qui ont été retrouvés.

En particulier, un fragment b_9 portant 3 hexoses a été détecté (flèche verte). Ce fragment correspond à une partie du peptide N-terminal précédemment identifié comme glycosylé avec un maximum de 4 hexoses. L'approche utilisée ici nous permet alors d'obtenir l'information complémentaire de la localisation de 3 de ces 4 hexoses sur les trois premières thréonines du peptide.

De plus, un fragment b_{75} portant 9 hexoses a également été identifié (flèche rouge). Etant donné que 4 hexoses au maximum sont portés par le peptide 1-27, une information supplémentaire est alors apportée : 5 hexoses seraient localisés sur le fragment de la protéine souligné en orange, portion de séquence non couverte lors de l'analyse des peptides tryptiques.

Les autres fragments glycosylés obtenus après fragmentation en source de toutes les protéoformes de la protéine, principalement N- ou C- terminaux, n'apportent pas d'informations supplémentaires quant à la localisation des hexoses sur la séquence en acides aminés.

L'information intéressante apportée ici par l'utilisation de ce mode de fragmentation est la présence de 5 hexoses sur la région de la protéine soulignée en orange. Au cours de l'analyse « bottom-up » des peptides tryptiques précédemment réalisée avec une fragmentation CID, un des deux peptides tryptiques pour lesquels aucun spectre MS/MS n'avait été obtenu se trouve dans cette région. Aucune information sur une éventuelle glycosylation de ce peptide d'intérêt n'avait alors été obtenue. De la même façon que pour le peptide C-terminal de la protéine, nous avons alors recherché, dans cette analyse « bottom-up » des peptides, les spectres MS dans lesquels sont retrouvés des pics ayant des masses correspondant à des valeurs de m/z théoriques du peptide d'intérêt. En revanche, contrairement au peptide C-terminal, nous n'avons pu mettre en évidence aucune glycosylation de ce peptide. Nous pouvons alors émettre une réserve quant au résultat apporté au cours de cette approche « top-down ».

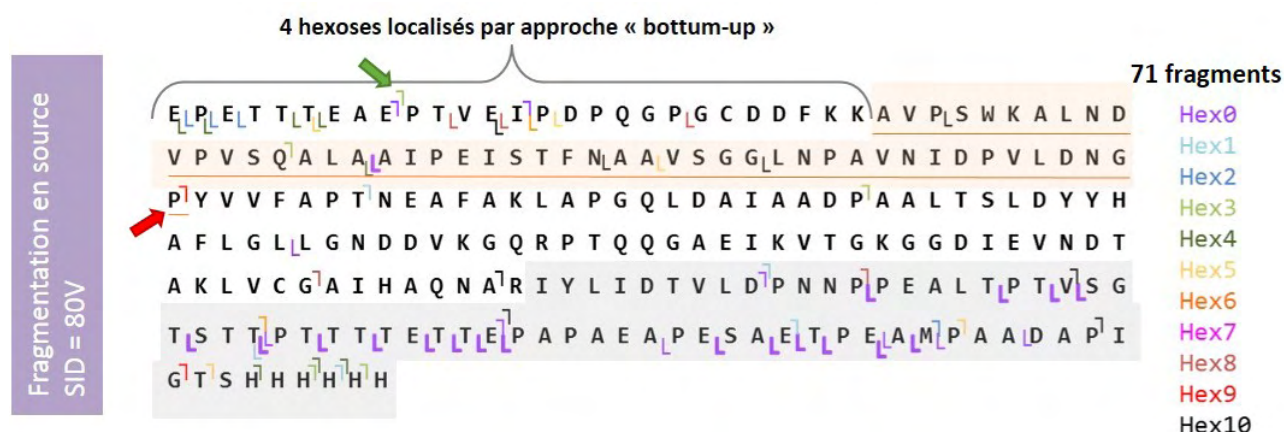


Figure 65: Les différents fragments obtenus après fragmentation de la protéine en source (fragmentation SID).

3. Résultats – fragmentation HCD

Avec une fragmentation HCD de la forme non glycosylée de la protéine, seulement 12 fragments (quasi-exclusivement des ions y) ont été retrouvés (figure 66). Ce mode de fragmentation utilisé sur les protéines est décrit dans la littérature comme conduisant à l'obtention d'un faible nombre de fragments, ce qui est en adéquation avec ces résultats. En revanche, après sélection des 4 premières formes glycosylées de la protéine (Hex1 à Hex5), ce sont les mêmes ions fragments qui sont obtenus, et aucun fragment glycosylé n'est retrouvé.

Il semblerait alors que la fragmentation HCD des protéines ne permet pas une localisation des hexoses et que ceux-ci soient éliminés des fragments, comme il est le cas lors de la fragmentation des peptides. De plus, ce mode de fragmentation utilisé sur les protéines semble peu informatif puisque peu de fragments sont obtenus.



Figure 66: Les différents fragments obtenus après fragmentation HCD de la protéine

4. Résultats – fragmentation ETD

Nous avons également testé une fragmentation ETD des protéines. En revanche, sur tous les spectres acquis au cours des différentes expériences avec ce mode de fragmentation, la résolution isotopique n'a pas été obtenue pour tous les ions au sein des spectres. En effet, des espèces aux charges réduites sont également détectées et sont très proches en masses des mêmes espèces sans charges réduites, le signal concernant les espèces d'intérêt est alors moins propre et la résolution isotopique n'apparaît pas. De plus, l'outil Xtract ne semble pas être l'outil le plus adapté pour la déconvolution de données issues de la fragmentation ETD, parmi lesquelles sont retrouvées des espèces non-radicales et des espèces radicales (Moradian, Kalli et al. 2014).

5. Conclusion

La fragmentation SID des protéines conduit à l'obtention d'un grand nombre de fragments, dont certains portent des hexoses. En revanche, étant donné qu'il n'y a pas de sélection d'un précurseur et que toutes les protéoformes présentes dans le spectromètre de masse peuvent subir la fragmentation, il est difficile d'interpréter les résultats obtenus. D'ailleurs, cette technique ne permet pas de confirmer ou d'infirmer avec certitude les résultats obtenus en « bottom-up », quant à l'absence de glycosylation dans la région de la protéine colorée en orange (figure 65).

La fragmentation HCD ne nous a pas permis d'aller plus loin dans la localisation des hexoses sur la protéine puisque aucun fragment glycosylé n'a été détecté.

Enfin, il n'a pas été possible de retraiter les données issues de la fragmentation ETD en raison de la qualité des spectres obtenus. Il serait alors nécessaire de revoir les paramètres d'acquisition utilisés pour cette fragmentation.

II. Recherche des sites de glycosylation via une approche de type « bottom-up » avec fragmentation ETD

1. Résultats – peptide N-terminal

Afin d'obtenir une localisation précise des sites de glycosylation de la protéine, nous avons réalisé une analyse de type « bottom-up » sur les peptides, avec une fragmentation ETD. Avant ce travail, de la même façon qu'avec l'analyse « bottom-up » des peptides tryptiques de la protéine avec la fragmentation CID, cette analyse avait été réalisée sur les peptides de la protéine non recombinante. De façon intéressante, une distribution non aléatoire des hexoses sur le peptide 1-26 avait été déterminée, avec des mannoses positionnés exclusivement sur les deux premières thréonines du peptide (Thr4 et Thr5).

RESULTATS troisième partie : Mise en place de méthodes d'analyse MS pour la caractérisation de la Fascicline

Afin de confirmer ces résultats, nous avons également analysé la digestion trypsique de la protéine recombinante, avec un mode de fragmentation ETD.

Des spectres MS/MS ont été obtenus pour les peptides 1-26 mono-, di- et tri- glycosylés (figure 67). De façon non ambiguë, il a été possible de localiser chacun des 3 hexoses sur le peptide : lorsque le peptide est mono-glycosylé, l'hexose est exclusivement positionné sur la thréonine 5 du peptide. Lorsque le peptide est di-glycosylé, chacune des thréonines 4 et 5 porte un hexose, tandis que lorsque le peptide est tri-glycosylé, un hexose est porté par la thréonine 4 et deux hexoses par la thréonine 5. Il semblerait alors qu'un certain déterminisme guide le processus de glycosylation sur le peptide N-terminal de la Fascicline.

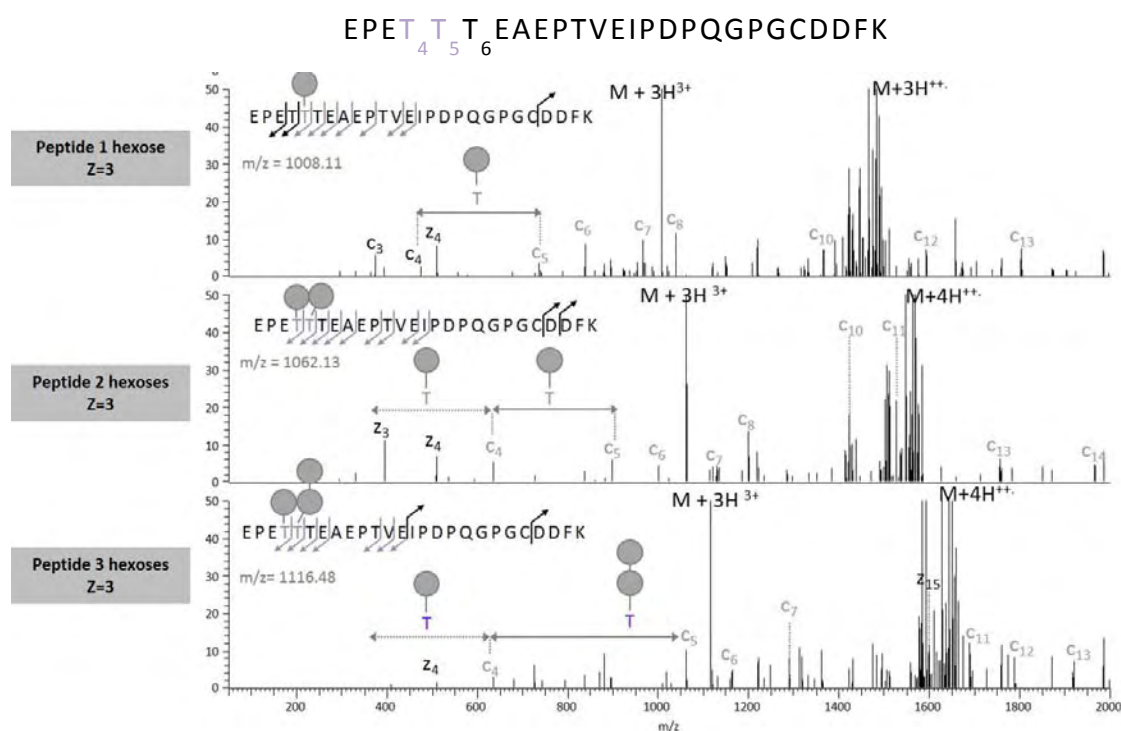


Figure 67: Spectres MS/MS avec fragmentation ETD des peptides N-terminaux mono-, di- et tri-glycosylés de la fascicline

En revanche, au cours de l'analyse réalisée sur les peptides de la protéine non recombinante, dans le cas du peptide N-terminal di-glycosylé, un hexose avait été positionné sur la thréonine 5, ce qui est en contradiction avec le résultat que nous avons ici obtenu. Nous pouvons émettre l'hypothèse que la localisation des hexoses est légèrement différente de la protéine non recombinante à la protéine recombinante qui est alors surexprimée et dont la séquence est modifiée du côté C-terminal. Il peut également être envisagé qu'au cours d'une telle analyse, basée sur l'interprétation de spectres MS/MS, il est possible de ne pas obtenir tous les spectres MS/MS permettant l'identification de toutes les combinaisons de positions de PTMs possibles. En effet, si l'on considère par exemple un peptide mono-glycosylé existant sous trois formes, les trois peptides ayant le même m/z, ils seront

sélectionnés et fragmentés de la même façon dans le spectromètre de masse. Etant donné qu'un ion d'un m/z donné n'est sélectionné qu'une fois pendant le temps de l'exclusion dynamique (voir introduction, partie A, paragraphe II.A.3.D.2)), il est possible que toutes les formes de l'ion ne soit pas toutes sélectionnées puis fragmentées.

2. Résultats – peptide C-terminal

Aucun spectre MS/MS n'a été obtenu au cours de cette analyse pour le peptide C-terminal, même si le mode ETD est pourtant plus adapté pour les peptides de hautes masses et de charges importantes (Swaney, McAlister et al. 2008). Afin de pouvoir aller plus loin dans l'étude de la glycosylation de la fascicline, et en particulier de ce peptide C-terminal, nous avons alors décidé de réaliser une double digestion trypsine/endoprotéinase V8 (digestion des protéines après les résidus acide aspartique et acide glutamique) afin d'obtenir des peptides dans la région C-terminale de masses et de charges plus faibles. Ainsi, il serait possible d'obtenir des spectres MS/MS couvrant la région C-terminale.

Quatre peptides doublement digérés à la trypsine et à l'endoprotéinase V8 contenus dans le peptide C-terminal tryptique ont été fragmentés (figure 68). Grâce aux spectres MS/MS ETD, il nous a été possible de localiser 6 hexoses sur 5 sites de glycosylation différents. De la même façon que pour le peptide N-terminal, il semblerait qu'il y ait un certain déterminisme dans la position des mannoses sur le peptide C-terminal (peptide 4 de la figure 67). Lorsque celui-ci est mono-glycosylé, le premier hexose est toujours positionné sur la thréonine 11, alors qu'un hexose sur cette même thréonine ainsi qu'un hexose sur la sérine 12 sont positionnés lorsque le peptide est di-glycosylé. Lorsque le peptide est tri-glycosylé, le troisième mannose vient alors se positionner sur la sérine 12.

Au cours de cette analyse et celle du peptide N-terminal, nous avons donc localisé 9 hexoses sur la protéine. Or, grâce à l'analyse de la protéine entière et également grâce au spectre MS déconvolué du peptide C-terminal, nous savons que la protéine porte une vingtaine d'hexoses. Il semblerait alors qu'il ne soit pas possible de localiser tous les hexoses sur la protéine avec les méthodes analytiques que nous avons utilisées jusque là.

Dans le cas où certaines thréonines ou sérines du peptide C-terminal seraient porteuses d'un long oligosaccharide de plus de 3 hexoses par exemple, nous pouvons imaginer que cette glycosylation gêne la digestion à la trypsine ou à l'endoprotéinase V8 et conduise à de nombreux clivages manqués et donc à des peptides de haut poids moléculaires pour lesquels les spectres MS/MS ne sont pas obtenus ou sont peu informatifs.

Nous avons alors entrepris de simplifier les motifs saccharidiques complexes éventuellement portés par certaines sérines ou thréonines en bloquant l'élongation des chaînes saccharidiques (Steentoft, Vakhrushev et al. 2011, Steentoft, Vakhrushev et al. 2013).

RESULTATS troisième partie : Mise en place de méthodes d'analyse MS pour la caractérisation de la Fascicline

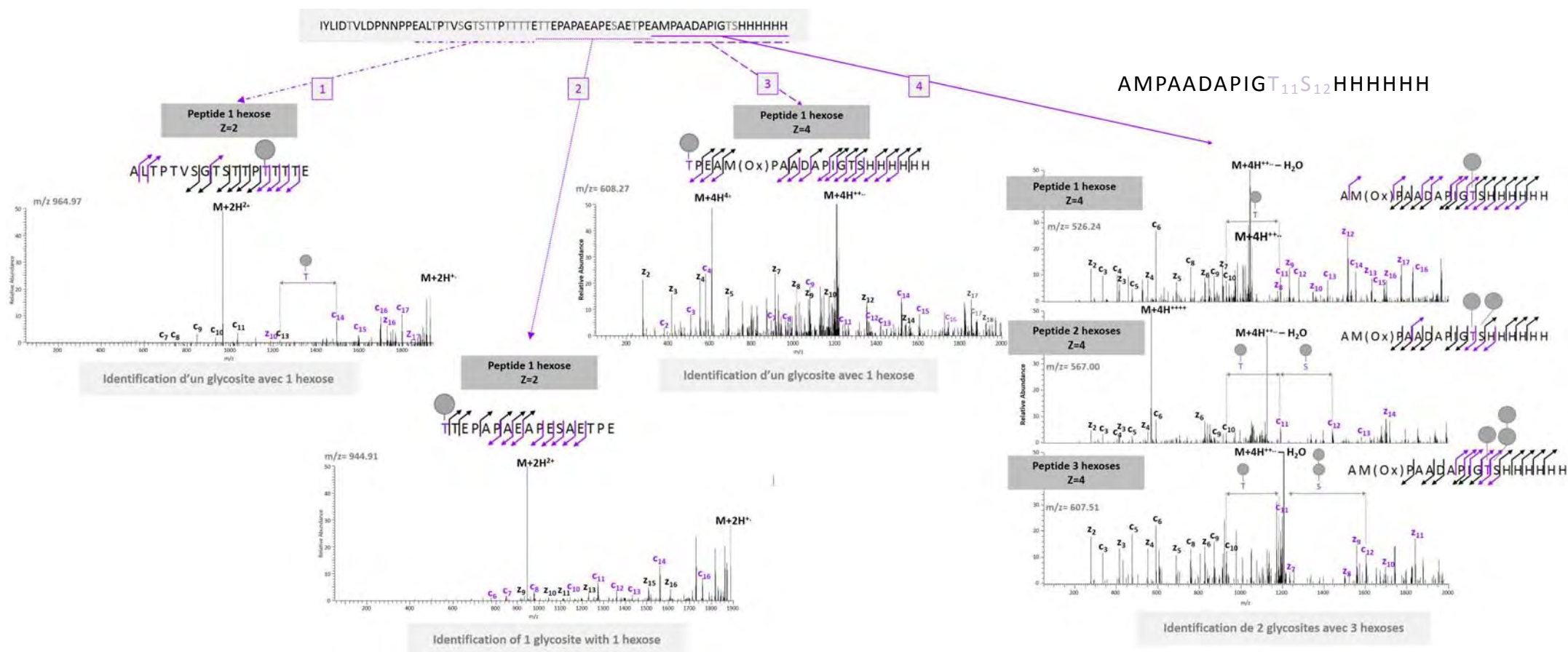


Figure 68: Spectres MS/MS des peptides de la région C-terminale de la Fascicline doublement digérés à la trypsine et à l'endoprotéinase V8.

III. Recherche des sites de glycosylation via une approche de type « bottom-up » sur des peptides aux motifs saccharidiques « simplifiés »

Nous savions que la protéine mannosyl-transferase codée par le gène *Msmeg_5447* est responsable de l'initiation de la *O*-mannosylation, donc du transfert du premier mannose sur une sérine ou une thréonine mais rien n'était connu sur l'enzyme responsable de l'élongation de la chaîne saccharidique. Or, la protéine PimE, codée par le gène *Msmeg_5149* est responsable de l'élongation des PIM5 en PIM6 (enchaînement α 1,2) (Hunter and Brennan 1990). La protéine codée par le gène *Msmeg_4247* est, elle, responsable de l'élongation des chaînes saccharidiques portées par les lipoglycanes (Chatterjee, Lowell et al. 1992). Ce sont les deux seules enzymes de *Mycobacterium smegmatis* du type α 1-2mannosyl transferases, c'est alors le rôle de chacune d'entre elle dans l'élongation des chaînes saccharidiques des protéines qui ont été évalués avant ce travail de thèse.

Les intensités de chacune des glycoformes du peptide N-terminal ont été recherchées parmi tous les peptides tryptiques des souches Δ *Msmeg_5149* et Δ *Msmeg_4247* (figure 69). Le profil des intensités de chacun des peptides N-terminaux (degrés de glycosylation de 0 à 3) dans la souche Δ *Msmeg_4247* semble être identique à celui de la souche sauvage, la protéine codée par le gène *Msmeg_4247* ne semble alors pas être responsable de l'élongation des chaînes saccharidiques des peptides. En revanche, le profil des intensités de ces mêmes peptides N-terminaux (degrés de glycosylation de 0 à 3) pour la souche Δ *Msmeg_5149* est différent, avec les formes non glycosylées et mono-glycosylées majoritaires, et un taux très faible de peptide di-glycosylé. Ceci qui suggère que c'est la protéine *Msmeg_5149* qui semble être responsable de l'élongation des chaînes saccharidiques des protéines.

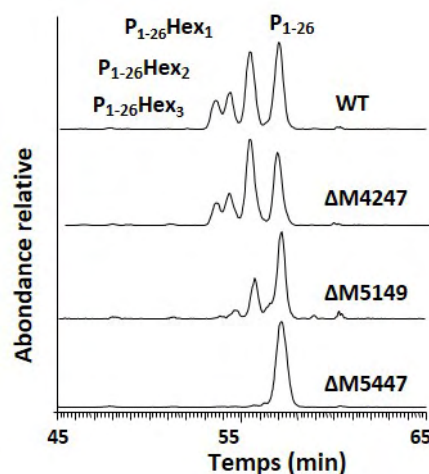


Figure 69: Profil des intensités des différentes glycoformes du peptide N-terminal dans les différentes souches sauvages, Δ *Msmeg_4247*, Δ *Msmeg_5149* et Δ *Msmeg_5447* (gène codant pour la PMT délété)

Nous avons donc évalué les intensités de chacune des formes glycosylées du peptide C-terminal, dans les mêmes souches (figure 70).

Par comparaison des intensités des différents peptides entre la souche sauvage et la souche Δ Msmeg_5149, la forme non glycosylée est plus représentée chez la souche Δ Msmeg_5149 que chez la souche sauvage. De même, les premières formes glycosylées du peptide semblent être plus intenses, tandis que les formes avec plus de 5 hexoses semblent être moins intenses. En revanche, le nombre d'hexose maximal sur ce peptide demeure important, à hauteur de 18 hexoses environ, contre 20 pour la protéine exprimée par la souche sauvage.

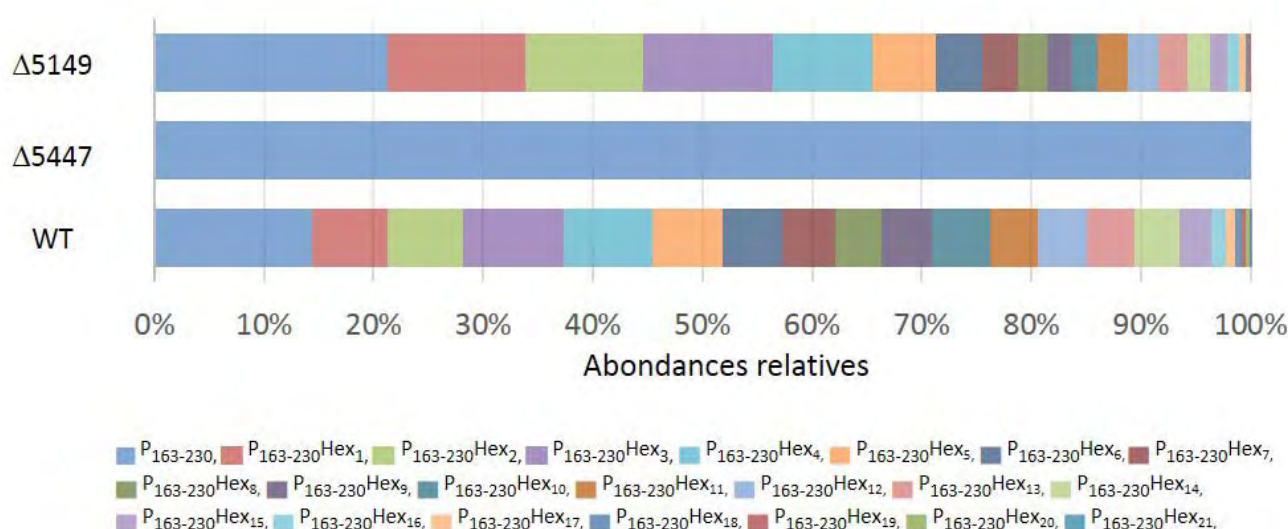


Figure 70: Abondances relatives de différentes formes glycosylées du peptide C-terminal de la fascicline

Nous pouvons alors faire l'hypothèse selon laquelle chaque hexose du peptide C-terminal est positionné sur une sérine ou thréonine, sans élongation des chaînes saccharidiques (il y a en effet une vingtaine de sérines et thréonines sur le peptide).

Une autre hypothèse que l'on peut avancer est l'hypothèse d'une autre enzyme qui prendrait le relais, l'enzyme codée par le gène Msmeg_4247 ou bien une autre enzyme non connue. Nous pourrions d'ailleurs construire un double mutant avec délétion des deux gènes Msmeg_4247 et Δ Msmeg_5149 pour évaluer ce point.

Afin d'essayer d'évaluer la première hypothèse, à savoir la présence de longues chaînes d'oligosaccharides dans la région C-terminale ou l'existence d'une seule forme de ce peptide avec un seul hexose par site de glycosylation potentiel (autant d'hexoses sont présents sur ce peptide qu'il n'y a de site potentiel de glycosylation), nous avons envisagé d'utiliser une technique de spectrométrie de masse innovante, la spectrométrie de masse de mobilité ionique. En effet, dans le cas d'un peptide avec présence de chaînes d'oligosaccharides, le

peptide existerait sous plusieurs formes : à des instants t_1 et t_2 , chaque chaîne d'oligosaccharides pourrait contenir un nombre d'hexoses différent (dynamique de l'élongation des chaînes), et pour un état de glycosylation global du peptide donné, plusieurs formes du peptide pourraient alors coexister. En revanche, dans le cas d'un peptide avec un seul hexose par site de glycosylation, ceci ne serait pas le cas. Nous allons voir que la spectrométrie de masse de mobilité ionique peut nous permettre de distinguer les deux cas.

IV. Recherche des sites de glycosylation par la technique de spectrométrie de masse de mobilité ionique

1. Principe

La mobilité ionique consiste à mesurer la vitesse d'un ion dans un gaz sous l'influence d'un champ électrique homogène (Uetrecht, Rose et al. 2010).

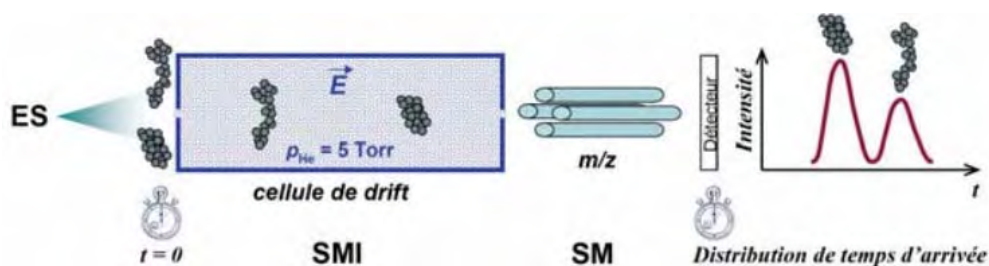


Figure 71: Principe de la spectrométrie de masse de mobilité ionique

A partir de la solution d'analytes en phase liquide, les ions sont produits par la source electrospray. Les ions sont ensuite injectés dans la cellule de mobilité ionique dans laquelle règne un champ électrique E (figure 71). Les ions entrent alors en collision avec les molécules de gaz présentes dans la cellule (sans subir de fragmentation). Un ion ayant une conformation étendue rencontrera d'avantage de molécules de gaz que le même ion avec une conformation plus compacte et sa vitesse de déplacement sera ralentie. L'ion à la conformation étendue arrivera alors plus tard au détecteur alors que l'ion à la conformation plus compacte y parviendra plus tôt (Gabelica 2006). Ainsi, la cellule de mobilité ionique permet la séparation de deux ions isobares possédant des conformations différentes.

C'est dans les années 1950-1960 que le premier tube de mobilité à champ électrique faible est apparu (McDaniel, Martin et al. 1962) . Il a fallu attendre les années 1997 pour voir apparaître la première analyse de petites protéines par couplage mobilité ionique – spectrométrie de masse (Clemmer and Jarrold 1997). Depuis, les techniques instrumentales ont été développées et les applications d'analyse de molécules biologiques par IM-MS se sont développées (Debaene, Wagner-Rousset et al. 2013).

Cette technique, puisqu'elle permet la séparation de deux peptides isobares aux conformations différentes, nous semblait alors particulièrement adaptée pour vérifier une des hypothèses données précédemment.

2. Résultats

En collaboration avec l'équipe de A. Van Dorsselaer et Sarah Cianferani du laboratoire de spectrométrie de masse bio-organique (LSMBO) à Strasbourg, nous avons voulu évaluer l'apport de la technique de séparation des ions par mobilité ionique pour la séparation de peptides glycosylés et la caractérisation des sites de glycosylation.

En effet, étant donné que la technique de mobilité ionique permet la séparation de deux espèces isobares possédant des conformations différentes, cette technique peut nous permettre de séparer deux peptides glycosylés isobares possédant le même nombre d'hexose(s) mais sur lesquels ces hexoses seraient positionnés sur des sites différents.

Nous avons tout d'abord testé l'apport de ce couplage sur les différentes formes du peptide N-terminal. A partir d'une digestion tryptique de la protéine, les différents peptides N-terminaux ont été séparés par HPLC. Une séparation complète de toutes les formes des peptides 1-26 et 1-27 n'a pas été possible étant donné que ces peptides sont tous élués à des temps de rétention très proches. 3 fractions ont alors été obtenues, avec une distribution des différents peptides différente d'une fraction à l'autre (les fractions ont été analysées et contrôlées sur un appareil de type MALDI-TOF) (figure 72). Ainsi, la première fraction contient par exemple les peptides 1-27 non glycosylé, mono, di- et tri-glycosylés, ainsi que le peptide 1-26 avec 3 ou 4 hexoses, en très faibles quantités. Dans cette fraction, c'est la forme tri-glycosylée du peptide 1-27 qui est majoritaire. Dans la deuxième fraction, la forme di-glycosylée du peptide 1-27 est la forme la plus abondante, et c'est la forme mono glycosylée de ce même peptide qui est majoritaire dans la fraction 3.

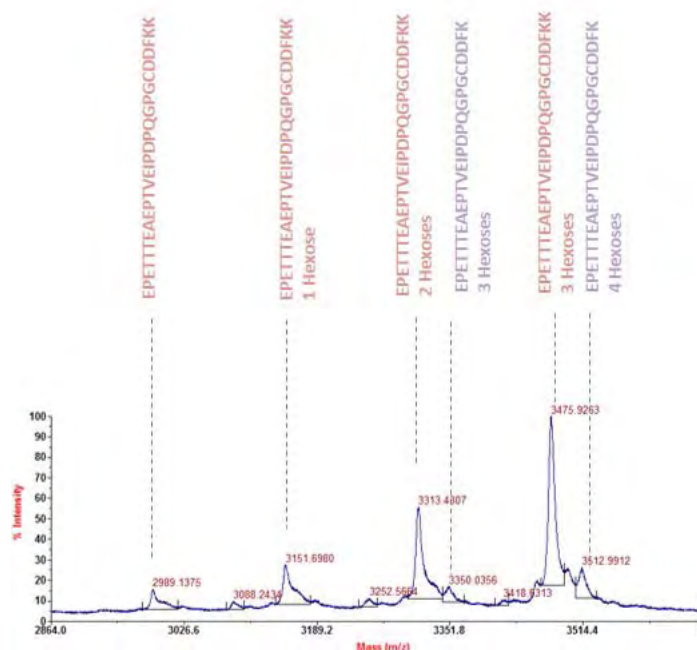
Les 3 fractions ont été analysées à Strasbourg par IM-MS sur un appareil Synapt G2 de Waters par Johann Stojko du LSMBO.

Après analyse des peptides de la fraction 1, un appel d'ions des 3 états de glycosylation du peptide 1-27, avec 3 ou 4 charges a été réalisé (figure 73). Des pics symétriques sont obtenus pour les 3 formes du peptide possédant 3 charges, suggérant une homogénéité conformationnelle de ces 3 peptides. En revanche, dans le cas des peptides avec 4 charges, parmi les pics issus des appels d'ions, un pic légèrement asymétrique est obtenu pour le peptide monoglycosylé. Cependant, cette asymétrie est relativement faible et peut ne pas être significative.

De la même façon, après analyse IM-MS des peptides de la fraction 2 et appels d'ions des ions d'intérêt, les pics obtenus correspondant aux peptides 1-27 avec 1, 2 ou 3 hexoses et 3 charges sont symétriques (figure 74). Cependant, un épaulement apparaît sur le pic correspondant au peptide di-glycosylé possédant 4 charges. Cet épaulement peut être le signe d'une hétérogénéité conformationnelle de ce peptide, donc de l'existence de deux isomères qui diffèrent par la position des deux hexoses sur la séquence en acides aminés. Le fait que cet épaulement soit visible sur le pic correspondant au peptide avec 4 charges et non sur le pic correspondant au même peptide avec 3 charges peut s'expliquer par le fait que le pouvoir résolutif de la cellule de mobilité ionique est plus important pour les ions de faibles m/z .

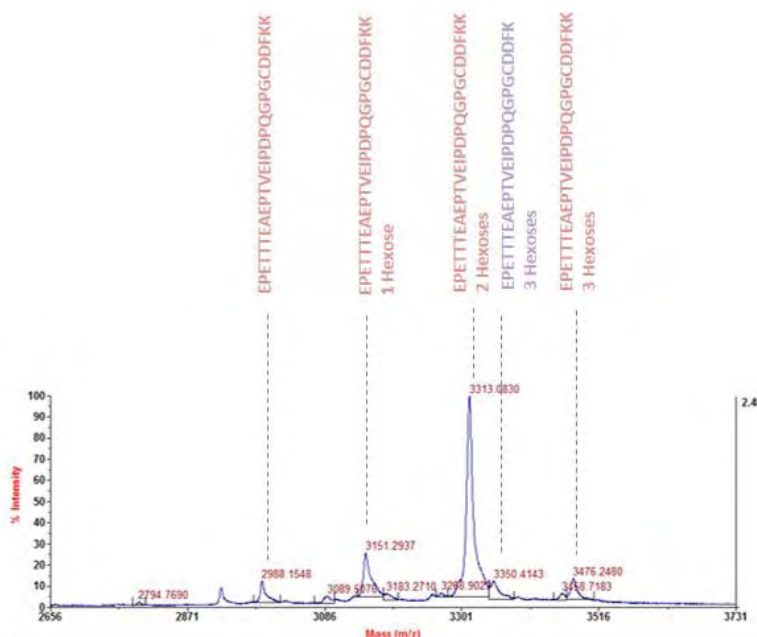
Ce même épaulement sur le même ion est retrouvé après analyse des peptides de la fraction 3, tandis qu'une légère asymétrie est visible sur le pic de l'ion monoglycosylé correspondant (figure 75).

Ces résultats semblent alors en faveur de l'hypothèse de deux isomères de position pour le peptide 1-27 di-glycosylé. Au cours des analyses réalisées dans le cadre de cette thèse sur la Fascicline recombinante, une seule forme du peptide avait été observée, avec un hexose sur chacune des thréonines 4 et 5. Or, au cours des analyses préliminaires réalisées sur la protéine non recombinante, une autre forme du peptide avec 2 hexoses sur la deuxième thréonine avait été observée, ce qui serait en faveur de ces résultats de spectrométrie de masse de mobilité ionique. Dans le cas où la forme du peptide di-glycosylé avec 2 hexoses sur la thréonine 5 serait en faible abondance par rapport à l'autre forme, comme il a déjà été évoqué, il serait possible que cette seconde forme n'ait pas été sélectionnée au cours de l'analyse bottom-up avec fragmentation ETD réalisée précédemment.



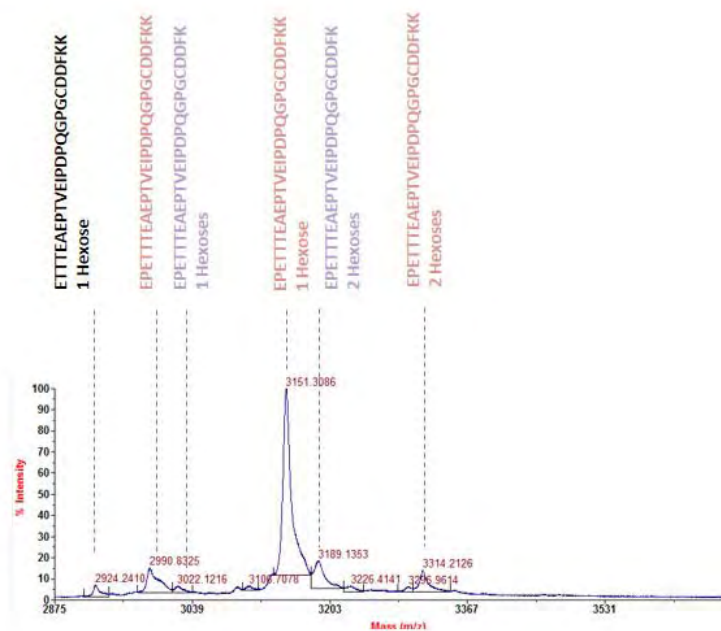
Fraction 1

Peptide 1-27 tri-glycosylé majoritaire



Fraction 2

Peptide 1-27 di-glycosylé majoritaire



Fraction 3

Peptide 1-27 monoglycosylé majoritaire

Figure 72: Spectres MALDI des fractions enrichies en différentes formes des peptides N-terminaux 1-26 et 1-27 de la fascicline

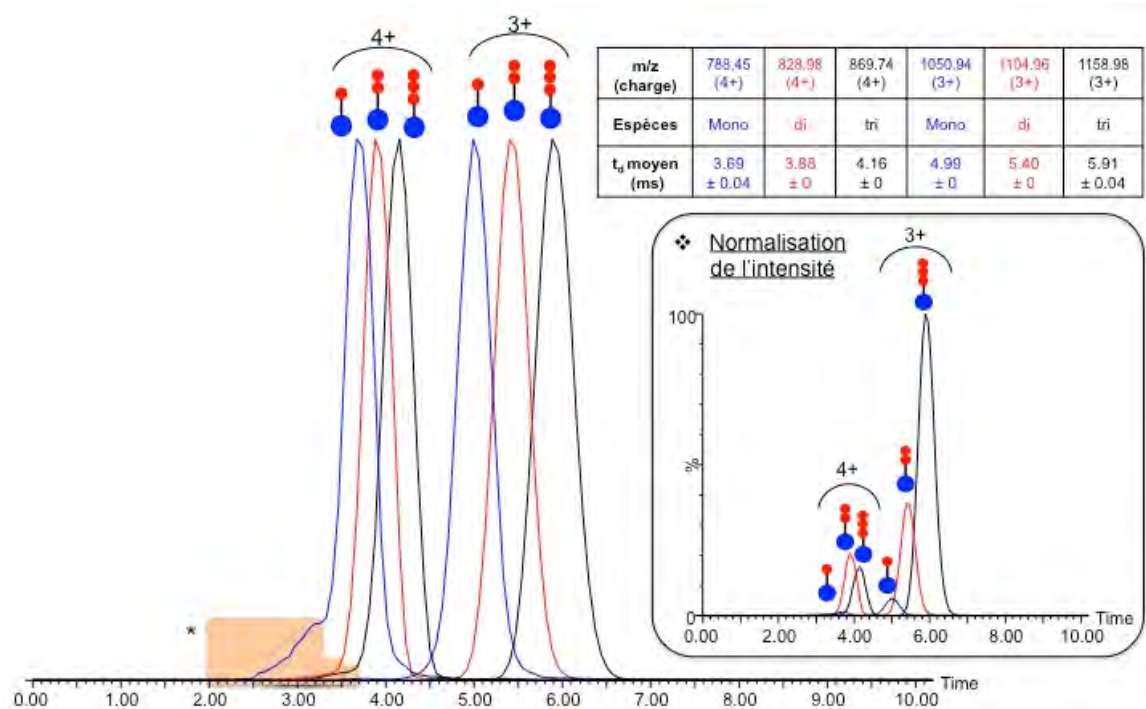
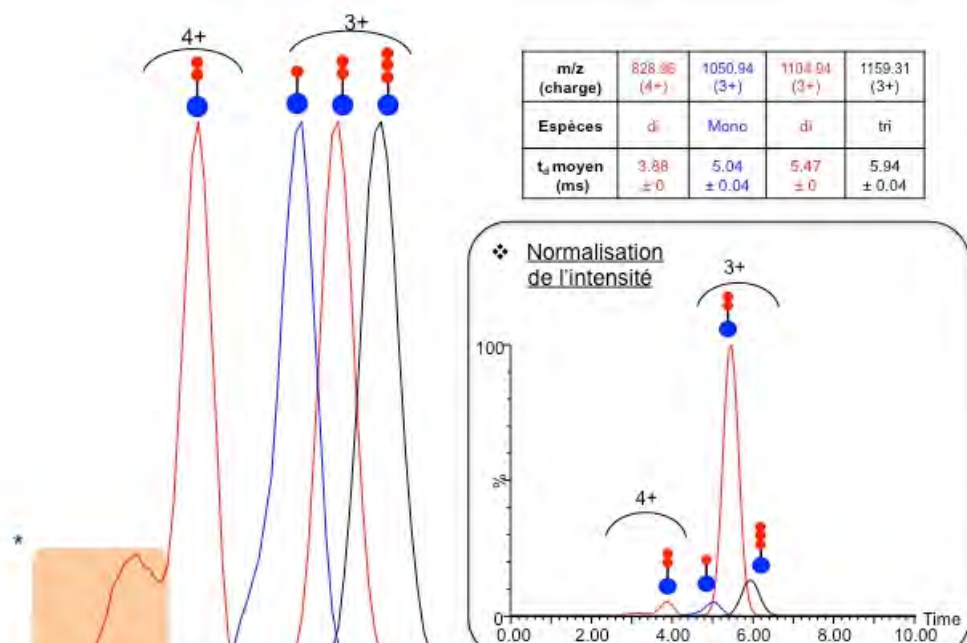


Figure 73: Spectre IMMS des différents peptides contenus dans la fraction 1



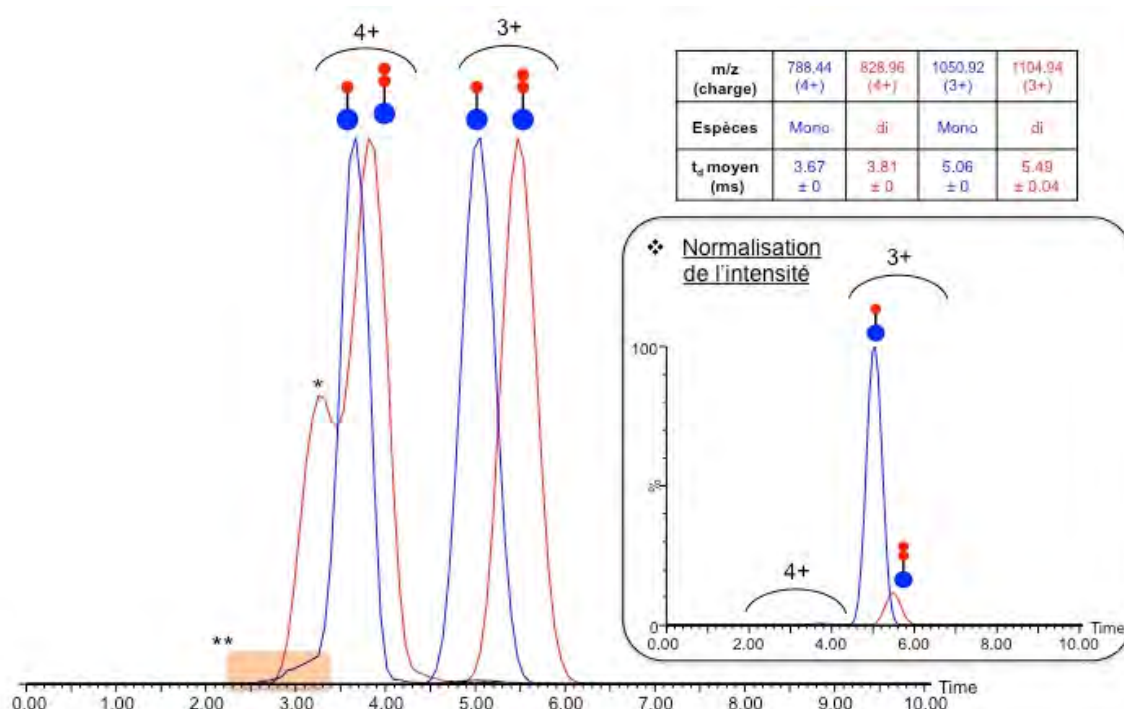


Figure 75: Spectre IMMS des différents peptides contenus dans la fraction 3

D. Conclusion

L'analyse de la protéine entière sur l'Orbitrap, avec une très bonne résolution, nous a permis d'obtenir rapidement le profil de la protéine et de déterminer le nombre d'hexoses portés par celle-ci.

Une analyse nanoLC-MS/MS des peptides tryptiques, avec un mode de fragmentation CID nous a ensuite permis d'identifier les peptides glycosylés : les peptides N-terminaux et C-terminaux. Des spectres MS/MS de bonne qualité ont été obtenus pour le peptide N-terminal, des spectres caractéristiques de peptides glycosylés, et un nombre maximum de 3 hexoses portés par ce peptide a été déterminé. En ce qui concerne le peptide C-terminal, malgré le fait qu'aucun spectre MS/MS n'a été obtenu, nous avons pu extraire des données le spectre MS de celui-ci et montrer la présence d'une vingtaine d'hexoses. Nous avons donc montré à ce stade de l'étude que c'est ce peptide C-terminal qui porte la majorité des hexoses.

Afin de parvenir à localiser les hexoses sur la protéine, et en particulier sur la partie C-terminale, nous avons tout d'abord envisagé d'utiliser la stratégie « top-down » sur la protéine entière, avec les fragmentations SID, HCD et ETD. Cependant, avec le mode de fragmentation ETD, qui apporte le plus grand nombre de fragments au cours d'une telle étude, la résolution isotopique n'a pas été obtenue pour les fragments et le retraitement des données n'a donc pas été possible. Ensuite, les fragmentations SID et HCD ne nous ont

pas permis d'obtenir des informations supplémentaires sur la localisation des hexoses sur la partie C-terminale, en raison de l'obtention de peu de fragments de faibles masses dans cette région.

Afin d'aller plus loin dans la caractérisation des glycosites, nous avons alors ensuite envisagé de réaliser l'analyse nanoLC-MS/MS des peptides tryptiques avec le mode de fragmentation ETD, adapté à l'analyse des PTMs labiles. Il nous a alors été possible de localiser les sites de glycosylation du peptide N-terminal. En revanche, même si le mode de fragmentation ETD est plus adapté pour la fragmentation de peptides de hauts poids moléculaires et de charges importantes, aucun spectre MS/MS n'a été obtenu pour le peptide C-terminal avec ce mode de fragmentation. Pour obtenir des peptides de masses moins importantes dans la région C-terminale, nous avons alors envisagé de réaliser une double digestion des peptides (trypsine/V8). L'analyse de ces peptides nous a alors permis de localiser 6 hexoses seulement sur 5 sites de glycosylation différents.

Il a été montré que la simplification de chaînes oligosaccharidiques en une seule unité osidique, et donc la diminution du nombre d'hexoses portés par un peptide, permettait une meilleure détermination des sites de glycosylation (Steentoft, Vakhrushev et al. 2013). Après avoir mis en évidence l'enzyme responsable de l'élongation des chaînes oligosaccharidiques, nous avons alors purifié la protéine produite chez un mutant de *Mycobacterium smegmatis* pour lequel le gène codant pour cette protéine avait été délété. Nous avons ensuite analysé les peptides de cette protéine par nanoLC-MS/MS. Cette technique ne nous a pas permis d'obtenir une meilleure caractérisation des sites de glycosylation de la protéine, étant donné qu'un nombre d'hexoses important était toujours présent sur le peptide C-terminal, en raison peut-être de l'existence d'une autre enzyme qui posséderait la même fonction que celle dont l'expression avait été bloquée. Nous pouvions également émettre l'hypothèse d'une élongation inexistante des motifs saccharidiques sur le peptide C-terminal et de la présence d'un hexose par site de glycosylation potentiel.

Afin de parvenir à déterminer l'existence du peptide C-terminal avec un seul hexose par site de glycosylation, ou bien avec la présence de chaînes oligosaccharidiques, nous avons utilisé la stratégie de spectrométrie de masse de mobilité ionique. Cette technique a tout d'abord été testée sur le peptide N-terminal, dont les sites de glycosylation étaient bien déterminés. Une ambiguïté existait concernant la position des hexoses sur le peptide 1-27 diglycosylé de la Fasciline : deux possibilités pour la position de ces hexoses avait été déterminé après analyse de la protéine non recombinante avant ce travail de thèse alors qu'une seule possibilité avait été déterminée au cours de ce travail (partie C.2). Cette dernière analyse par spectrométrie de masse de mobilité ionique a confirmé cette ambiguïté puisque une hétérogénéité conformationnelle a été suggérée pour ce peptide di-glycosylé. En revanche, la fraction contenant les différentes formes du peptide C-terminal était trop complexe pour pouvoir être analysée par cette technique.

Avec l'utilisation de plusieurs outils de spectrométrie de masse, nous avons donc tenté d'obtenir une caractérisation complète de la glycosylation de la fasciline. En revanche, nous avons été confronté à certaines limites inhérentes à l'utilisation de ces techniques.

En effet, comme déjà évoqué, au cours de toute analyse des peptides par nanoLC-MS/MS, il est possible que les spectres obtenus ne soient pas représentatifs de toutes les formes des peptides et que les informations obtenues quant à la localisation des PTMs soient incomplètes.

De plus, au cours d'une analyse d'une protéine par « top-down », afin d'obtenir suffisamment d'informations sur la localisation des PTMs, il est nécessaire d'avoir une couverture de séquence suffisante. Au cours des analyses que nous avons réalisées, les couvertures de séquences obtenues ont été trop faibles pour obtenir suffisamment d'informations sur les modifications étudiées. De plus, il serait nécessaire de réaliser un fractionnement des protéines en amont de prochaines analyses pour séparer les différents états de glycosylation de la protéine et ainsi essayer de décomplexifier l'étude.

De même, aucun résultat n'a été obtenu en spectrométrie de masse de mobilité ionique pour le peptide C-terminal en raison d'échantillons trop complexes.

CONCLUSION GENERALE ET PERSPECTIVES

Lors de mon arrivé au laboratoire il a pu être mis en évidence le caractère indispensable de la *O*-mannosylation des protéines bactériennes pour la virulence de *Mycobacterium tuberculosis* confirmant ainsi la contribution majeure de mannoprotéines du bacille dans la physiopathologie de l'infection. Au cours de cette thèse, j'ai donc mis en place diverses méthodes de protéomique et spectrométrie de masse pour identifier et caractériser ces mannoprotéines afin de mieux comprendre leurs rôles dans la virulence du bacille tuberculeux.

Dans un premier axe, j'ai mis en place une stratégie glycoprotéomique pour une identification des glycoprotéines du bacille. Pour cela, l'étude a débuté par l'analyse nanoLC-MS/MS des protéines sécrétées par la bactérie dans un milieu de culture 7H9 ADC et la mise en place d'un outil bioinformatique dédié pour la recherche des données d'intérêt parmi tout le jeu de données acquises. Cet outil bioinformatique repose sur la recherche de spectres MS/MS caractéristiques des glycopeptides parmi tous les spectres MS/MS enregistrés. Une première version de l'utilisation de cet outil bioinformatique appliqué à l'analyse des protéines sécrétées par la souche sauvage dans ce milieu de culture nous a permis d'identifier une liste de quelques glycoprotéines candidates possédant de faibles probabilités de glycosylation et dont ne faisaient pas partie les 4 glycoprotéines connues à ce jour.

Nous avons alors essayé d'évaluer l'apport de l'utilisation de méthodes d'enrichissement dans le cadre de notre étude, afin d'améliorer ce résultat. Nous avons utilisé des méthodes d'enrichissement par affinité à la concanavaleine A au niveau protéique et au niveau peptidique, ainsi que l'apport de l'enrichissement des peptides sur phase HILIC. L'enrichissement par affinité à la concanavaleine A au niveau protéique n'a pas mené à des résultats intéressants, contrairement au même enrichissement au niveau peptidique. En revanche, par manque de reproductibilité de ces résultats au niveau peptidique, nous avons choisi de ne pas poursuivre l'utilisation de cette méthode. L'essai d'enrichissement des peptides sur phase HILIC a souffert du manque de spécificité de cette méthode pour les peptides glycosylés et n'a permis l'identification d'aucune glycoprotéine. Pour la suite de l'étude, nous avons alors décidé de ne pas poursuivre l'utilisation de ces méthodes d'enrichissement pré-analytiques.

En revanche, un des moyens de contrôler la validité de la première liste de glycoprotéines était de rechercher les glycoprotéines de la souche mutante $\Delta Rv1002c$ invalidée pour la Protein Mannosyl Transferase et de vérifier, chez cette souche mutante, l'absence de glycosylation des glycoprotéines identifiées chez la souche sauvage. Or, la souche mutante présentait un défaut de croissance dans le milieu 7H9 ADC. Nous avons alors entrepris d'identifier les glycoprotéines sécrétées dans un autre milieu de culture, le milieu Sauton. L'étude a donc tout d'abord été poursuivie avec l'analyse nanoLC-MS/MS de deux réplicats

de protéines sécrétées par la souche sauvage dans un milieu de culture Sauton. L'utilisation de l'outil bioinformatique dédié a été modifiée (introduction de filtres supplémentaires) pour une identification des glycoprotéines avec un meilleur indice de confiance. Dix glycoprotéines possédant de bonnes probabilités de glycosylation ont alors été identifiées, dont une des glycoprotéines connues. La même stratégie appliquée à la recherche des glycoprotéines de la souche complémentée $\Delta Rv1002c$: $\Delta Rv1002c$ a conduit à l'identification de 11 glycoprotéines supplémentaires. Par la suite, comme évoqué précédemment, nous avons également appliqué la même stratégie de recherche des glycoprotéines chez la souche mutante et nous avons bien vérifié l'absence de glycosylation des glycoprotéines identifiées jusque-là chez la souche sauvage et chez la souche complémentée.

Enfin, une fois la stratégie de recherche des glycoprotéines mise en place, celle-ci a été appliquée à la recherche des glycoprotéines présentes dans les extraits cellulaires de la souche sauvage et de la souche complémentée du bacille. Ce sont 4 glycoprotéines supplémentaires qui ont été validées.

Suite à l'analyse de protéines sécrétées et issues de la lyse des bactéries des souches sauvage et complémentée de *Mycobacterium tuberculosis* cultivées en milieu Sauton, nous avons donc identifié 25 glycoprotéines avec un fort indice de confiance, dont 10 nouvelles¹, 5² (dont LpqH) prédites (sur la base de leur interaction avec la con-A) pour lesquelles nous avons pu identifier des peptides glycosylés et 10 (dont Apa et Lprf) des 16 glycoprotéines mycobactériennes caractérisées à ce jour (62%). Parmi ces protéines, certaines sont connues pour avoir un rôle dans la virulence du bacille et constituent des candidats intéressants pour tenter d'analyser plus avant l'impact de la glycosylation sur leurs propriétés biologiques.

Dans un second axe, afin d'évaluer l'effet systémique de l'arrêt du processus de glycosylation sur l'expression globale des protéines, nous avons entrepris de réaliser une analyse protéomique quantitative des protéines sécrétées et issues de la lyse des bactéries de la souche mutante, de la souche sauvage et de la souche complémentée. Nous avons comparé les expressions des protéines à deux temps de culture, un temps de culture correspondant à la phase exponentielle de croissance et un temps correspondant à la phase stationnaire. Les résultats préliminaires de ces études font apparaître des différentiels d'expressions de certaines protéines entre les souches sauvage et mutante qui pourraient être spécifiquement associés à l'invalidation du gène *Rv1002c* de la protéine-O-Mannosyl-transferase. Ces données encore en cours d'analyse seront confrontées par aux résultats de l'analyse transcriptomique réalisée sur les mêmes échantillons dans l'objectif de préciser l'impact systémique de l'interruption du processus de O-mannosylation sur la biologie du bacille et d'essayer de mettre en évidence des altérations métaboliques ou physiologiques pouvant rendre compte de l'atténuation de la virulence.

¹ Rv0822, Rv0838, Rv1743, Rv1754, Rv2969, Rv3584, Rv3787, Rv3623, Rv3705a, Rv3706c,

² Rv0315, Rv0928, Rv1411, Rv3244, Rv3763/19 kDa lipoprotein antigen LPQH,

Enfin, dans un troisième axe, nous avons mis en place une stratégie d'analyse complète des glycoprotéines mycobactériennes, dans le but de pouvoir utiliser cette stratégie pour l'analyse de glycoprotéines d'intérêt de *Mycobacterium tuberculosis*. Pour cela, nous avons travaillé avec une glycoprotéine modèle de *Mycobacterium smegmatis* identifiée au laboratoire, la Fasciline.

En effet l'analyse de la protéine entière a permis de mettre en évidence un profil de glycosylation inattendu avec un nombre d'hexose totalement atypique supérieur à 20 unités.

Ceci nous a conduit à analyser le digest tryptique de la protéine par nanoLC-MS/MS avec fragmentation CID afin d'identifier les peptides glycosylés de la protéine. Des spectres caractéristiques de peptides glycosylés ont été obtenus pour le peptide N-terminal (jusqu'à 3 hexoses). En revanche, en raison de sa masse importante, aucun spectre MS/MS n'a été obtenu pour le peptide C-terminal.

Nous avons alors entrepris de réaliser une double digestion à la trypsine et à l'endoprotéinase V8 de la protéine pour réduire la taille des peptides issus de l'extrémité C-terminale. Après analyse nanoLC-MS/MS avec fragmentation ETD des peptides tryptiques et des peptides doublement digérés, ce sont 3 hexoses sur le peptide N-terminal et 6 hexoses dans la région C-terminale qui ont été localisés.

Afin d'aller plus loin dans la caractérisation, alors incomplète, des sites de glycosylation de la protéine, nous avons essayé de mettre en place une méthode « top-down » d'analyse de protéines. Toutefois, après fragmentation SID, HCD et ETD de la protéine, aucune indication supplémentaire sur la localisation des sites de glycosylation n'a été apportée par cette méthode en raison d'une couverture de séquence de la protéine insuffisante.

Enfin, afin de mieux comprendre la répartition des hexoses dans la région C-terminale de la protéine, à savoir par exemple si la glycosylation consiste principalement en une unité osidique par site de glycosylation ou si ce sont principalement des chaîne d'oligosaccharides que nous retrouvons dans cette région, nous avons essayé d'évaluer l'apport de la spectrométrie de masse à mobilité ionique pour répondre à cette question, en collaboration avec le laboratoire de spectrométrie de masse bio-organique de Strasbourg. Cette technique nous a permis de confirmer la répartition des hexoses sur le peptide N-terminal précédemment déterminée grâce à l'analyse nanoLC-MS/MS avec fragmentation ETD des peptides. En revanche, le mélange peptidique des différentes formes du peptide C-terminal était trop complexe pour pouvoir être analysée par cette technique.

Grâce à l'outil bioinformatique « Neutral Loss » développé au laboratoire et dédié à la recherche de protéines possédant des PTMs labiles, nous avons donc réussi à identifier 25 glycoprotéines de *Mycobacterium tuberculosis*, avec pour chacune d'entre elles le ou les peptide(s) glycosylé(s) et le nombre d'unités hexoses. De plus, nous avons vérifié la fiabilité des résultats obtenus en montrant, chez la souche mutante dépourvue de glycosylation des protéines, l'absence de glycosylation de toutes les glycoprotéines identifiées chez la souche sauvage et chez la souche complémentée. De plus, ce résultat a été obtenu sans utiliser de

méthodes d'enrichissement. Pour compléter cette liste, nous pourrions alors envisager dans un premier temps de réaliser l'analyse des protéines de réplicats supplémentaires de la souche sauvage du bacille. Dans un second temps, il serait également envisageable de tester et d'utiliser d'autres méthodes d'enrichissement que celles que nous avons essayé d'utiliser au cours de cette thèse.

La littérature indique que certaines des glycoprotéines que nous avons identifiées seraient impliquées dans la virulence du bacille. Afin d'évaluer le rôle de chacune des glycoprotéines identifiées dans le phénotype observé pour le mutant $\Delta Rv1002c$, il a été décidé de construire les mutants de certaines d'entre elles, choisies selon les données bibliographiques (construction des mutants avec délétion du gène de chacune des glycoprotéines). Cette étude s'inscrit dans le cadre du projet ANR GlycoPrOmicS. Le phénotype de chacun de ces mutants sera étudié et comparé au phénotype du mutant $\Delta Rv1002c$, *in vivo* et *in vitro* chez la souris. Pour tous les mutants présentant des caractéristiques phénotypiques comparables à celles du mutant $\Delta Rv1002c$, une étude approfondie de la protéine d'intérêt sera menée afin de préciser les sites de glycosylations de la protéine. Il sera alors possible de déterminer formellement l'implication des motifs glycaniques dans les propriétés fonctionnelles de ces glycoprotéines en abrogeant sélectivement la glycosylation sur chacun de ces sites par mutation ponctuelles.

Enfin, dans le but de proposer une stratégie de caractérisation complète des glycoprotéines, et en particulier permettant de caractériser les sites de glycosylation, nous avons évalué au cours de cette thèse nombre d'approches couramment utilisées pour l'analyse des glycoprotéines par spectrométrie de masse. Ainsi, nous avons prospecté les méthodes d'analyse de type « Top-Down » des protéines entières pour tenter de déterminer plus précisément le profil de glycosylation de la Fasciline. Toutefois ces approches se sont avérées encore insuffisamment résolutive pour pouvoir aborder la complexité du mélange de glycoformes et protéoformes résultant de l'hétérogénéité intrinsèque du processus de glycosylation. Ensuite, nous avons exploré les possibilités offertes par les approches « bottom-up » avec la fragmentation ETD pour la caractérisation des glycosites. Bien que confrontés aux limites de cette approche, nous sommes parvenus à caractériser environ 50% des glycosites de la protéine. Par ailleurs, l'approche de spectrométrie de masse par mobilité ionique s'est montrée efficace pour la détermination du nombre de formes différentes du peptide N-terminal, incluant les isomères de positions des motifs hexoses. Toutefois, ici encore c'est la complexité de l'échantillon contenant les différentes formes du peptide tryptique C-terminal qui a été limitante. Pour obtenir une caractérisation plus complète des glycosites de la protéine, l'effort doit alors se porter sur la préparation des échantillons en amont de l'analyse. En effet, le peptide tryptique C-terminal est par exemple présent avec plus de 20 degrés de glycosylation différents et chacun de ces états de glycosylation comporte plusieurs molécules isobares se différenciant par la structure et la position des motifs

glycaniques sur la chaîne peptidique Analyser toutes ces formes au sein d'un même échantillon demeure un défi analytique avec les techniques actuelles disponibles et nécessite impérativement de simplifier au mieux l'échantillon de départ si l'on veut avoir accès à l'étendue de la complexité de ces molécules.

MATERIEL ET METHODES

I. Préparation des échantillons pour l'analyse par spectrométrie de masse: digestion en gel

Les échantillons d'intérêt sont déposés sur un gel SDS-PAGE. La séparation est réalisée dans gel de polyacrylamide à 12%, dans du tampon de migration TGS (Tris Glycine SDS) 1X (Biorad), à un voltage de 80V par gel. Dans certains cas, l'analyse est arrêtée lorsque les protéines sortent du gel de concentration (« stacking ») et dans d'autres cas, les protéines sont séparées dans le gel de séparation jusqu'à ce que le front de migration atteigne le bas du gel. Parallèlement aux échantillons, un marqueur de poids moléculaire (Kaléidoscope, Biorad) est également déposé sur le gel.

Le gel est ensuite mis à colorer au minimum 15 minutes dans une solution de bleu de coomassie colloïdal (Instant Blue, Fermentas) puis rincé à l'eau. Enfin, le gel est scanné (Molecular Imager GS-800 Calibrated Densitometer, Biorad) puis les pistes des échantillons sont découpées en un nombre de bandes défini. Dans le cas où la migration a été arrêtée juste après l'entrée des protéines dans le gel de concentration, une seule bande contenant l'ensemble de l'échantillon est découpée. Chaque bande est découpée en petits morceaux de gels placés dans des tubes contenant de l'eau milliQ.

Lavage des morceaux de gel

Les morceaux de gels sont rincés à l'eau milliQ puis déshydratés dans un mélange acetonitrile (ACN) /50 mM bicarbonate d'ammonium (50/50, v/v). Ils sont ensuite mis sous agitation 15 min à 37°C. Ces étapes sont répétées jusqu'à ce que les morceaux de gel aient perdu leur coloration bleue, puis le gel est déshydraté à l'ACN pur. Ils sont ensuite séchés au speed-vac.

Digestion trypsique

Les protéines sont digérées *in-gel* en plaçant les morceaux de gel dans une solution de trypsine (ratio enzyme/substrat d'environ 1/50, p/p) dans du tampon bicarbonate d'ammonium 50 mM. Une réhydratation des bouts de gel est d'abord effectuée sur glace avec la solution de trypsine pour que l'enzyme pénètre bien tout le volume du gel tout en restant inactive. Du tampon bicarbonate d'ammonium 50 mM est ajouté au besoin pour bien immerger les bouts de gel. La digestion s'effectue ensuite pendant 12H à 37°C.

Double digestion endoprotéinase V8 – trypsine

Les protéines sont digérées *in-gel* en plaçant les morceaux de gel dans une solution d'endoprotéinase V8 (ratio enzyme/substrat d'environ 1/40, p/p) dans du tampon phosphate 25 mM pH 7,8. La digestion s'effectue ensuite pendant 12H à 37°C. Le mélange est ensuite chauffé à 95°C pendant 10 minutes afin de désactiver l'endoprotéinase V8. A la solution d'endoprotéinase et de tampon phosphate est ajoutée une solution de trypsine (ratio enzyme/substrat d'environ 1/20, p/p) dans du tampon phosphate. La seconde digestion s'effectue également pendant 12H à 37°C.

Extraction peptidique

Les peptides tryptiques sont extraits par des incubations puis sonications successives, une première fois dans du bicarbonate d'ammonium 50 mM, puis deux fois dans un mélange acétonitrile/acide formique (1:1). Après chaque sonication, les surnageants sont récupérés et rassemblés pour constituer l'extrait peptidique.

Les extraits ainsi obtenus sont enfin évaporés au speed-vac puis conservés à -20°C.

II. Analyse nanoLC-MS/MS et retraitement des données de glycoprotéomique et de protéomique quantitative

Analyse nanoLC-MS/MS :

L'analyse protéomique des échantillons est réalisée par nano-LC-MS/MS grâce à un système de nano-HPLC (U3000, Dionex) couplé à un spectromètre de masse (LTQ-Orbitrap Velos avec module ETD, Thermo Fischer Scientific). Les extraits peptidiques secs sont repris dans 14 µL de solvant A' (5% ACN, 0.05% acide trifluoro acétique (TFA)) et 5 µL sont injectés sur une pré-colonne de phase inverse C18 (300 µm ID x 5 mm, Dionex) à un débit de 20 µL/min. Après 5min de dessalage, la pré-colonne est basculée en ligne avec une colonne analytique de phase inverse C18 (75 µm ID x 15 cm PepMap C18, Dionex) équilibrée avec 95 % de solvant A (5% ACN, 0,2 % FA) et 5% de solvant B (80% ACN, 0,2% FA). L'élution des peptides est réalisée par un gradient croissant de 5% à 50 % de solvant B sur 160 min (**pour les analyses de glycoprotéomique**) ou 8H (**pour les analyses de protéomique quantitative**) à un débit de 300 nL/min. L'acquisition est réalisée en mode « Data Dependant Acquisition » et est pilotée par le logiciel Xcalibur (Thermo Fischer Scientific). Les scans MS sont enregistrés dans l'analyseur Orbitrap avec une résolution de 60000 à m/z=400. Les 20 ions les plus intenses de chaque scan MS sont sélectionnés pour subir en parallèle une analyse

MS/MS dans la trappe linéaire LTQ (fragmentation CID). Un temps d'exclusion dynamique de 60 sec a été utilisé.

Pour les analyses avec fragmentation ETD, un temps d'activation de 150 ms a été utilisé.

Recherche dans les bases de données et validation :

Les recherches en banques de données ont été réalisées avec le logiciel Mascot Daemon (version 2.2.0, Matrix Science) en utilisant la macro Extract_msn.exe de Xcalibur (Thermo Fisher Scientific) pour générer les peaklists. L'interrogation a été réalisée pour l'espèce *Mycobacterium tuberculosis* dans la banque de données Uniprot en incluant des modifications fixes (carbamidométhylation des cystéines) et variables (oxydation des méthionines, glycosylation par des motifs hexoses sur les sérines et thréonines, déamidation des asparagines).

Les tolérances de masse en mode MS et MS/MS ont été respectivement fixées à 8 ppm et 0,6 Da. La recherche a été effectuée avec la possibilité de deux clivages manqués et pour des peptides de charges 2+ et 3+. Les protéines identifiées ont ensuite été validées automatiquement avec le logiciel MFPaQ développé au laboratoire pour obtenir un taux de faux-positifs inférieur ou égal à 1% sur les peptides et les protéines.

Retraitement des données de glycoprotéomique

A l'issue de la recherche dans les bases de données, les « peak list » créés par Mascot sont entrés dans le logiciel Neutral Loss. Le logiciel sélectionne comme spectres de glycopeptides putatifs les spectres dans lesquels un ou des ions possèdent une masse proche de la masse théorique d'un ou de plusieurs ions provenant de la perte d'un ou de plusieurs hexoses à partir de l'ion parent, avec une tolérance de masse fixée à 500 ppm au début de l'étude puis à 800 ppm à la fin de l'étude. De plus, pour ces ions d'intérêt, un rang maximal au sein de spectres de 15 a été fixé, ainsi qu'une intensité relative minimale de 20%.

Quantification avec MFPaQ pour les données de protéomique quantitative :

Une analyse quantitative différentielle des différentes conditions a été réalisée avec le module de quantification sans marquage de MfPaQ. A partir des résultats d'identification validés, les profils chromatographiques des peptides correspondants sont recherchés dans les fichiers bruts à comparer. Le logiciel apparie dans un premier temps les peptides identifiés validés dans toutes les acquisitions MS, puis les utilise pour aligner ces acquisitions MS en temps de rétention. A partir de la matrice d'alignement en temps de rétention obtenu, le logiciel prédit le temps de rétention des ions peptidiques dans les analyses où ils n'ont pas été identifiés par MS/MS, et extrait le signal correspondant à partir du temps de

rétenion prédit, et de la masse exacte du peptide. Les aires des profils chromatographiques des peptides identifiés sont ainsi extraits dans chacun des fichiers bruts et sont utilisés pour réaliser la quantification de chaque peptide. MFPaQ calcule ensuite pour chaque protéine un indice d'abondance protéique, le PAI (« Protein Abundance Index »), qui correspond à la moyenne des intensités des trois peptides les plus intenses d'une protéine (définis sur l'ensemble des conditions comparées, et identiques entre ces différentes conditions). Ces PAIs sont ensuite utilisés pour calculer le ratio d'abondance relative d'une même protéine dans les différentes conditions.

Analyse des données avec le logiciel STRING: Afin de visualiser les protéines différentiellement exprimées possédant une interaction dans les différentes conditions, les numéros d'accès UniProt des protéines ont été entrés dans le logiciel STRING (<http://string-db.org/>; Snel et al 2000).

III. Analyse ESI-MALDI des peptides

10 µg de peptides digérés à la trypsine ou digérés à l'endoprotéinase V8 et à la trypsine en solution sont déposés sur une plaque MALDI. Lorsque le dépôt est sec, la matrice acide α -cyano-4-hydroxycinnamique est ensuite déposée.

L'analyse MS des échantillons est ensuite réalisée sur un appareil Maldi ToF-ToF 4700 (Applied biosystems) en mode réflectron. L'énergie du LASER utilisée est de 4000 kW.cm⁻².

IV. Enrichissement en glycoprotéines ou glycopeptides

Enrichissement sur Concanavalline A au niveau protéique

1 mg de protéines du surnageant de culture de *Mycobacterium tuberculosis* en milieu Sauton ont été précipités à l'acétone froid. 1,5 mL de phase de concanavalline A greffée sur des billes d'agarose (Sigma) sont prélevés et lavés 3 fois avec 8 mL de tampon de lavage Tris.HCl 20 mM, NaCl 0.15 M, MnCl₂ 1 mM, CaCl₂ 1 mM à pH 7,4. Les protéines précipitées sont mises en solution avec 20 mL de tampon de lavage puis mises en contact avec la phase. Le mélange est incubé à 4°C pendant 2H. A l'issue de l'incubation, le mélange est centrifugé à 5000g pendant 15 minutes. Le surnageant est conservé et la phase est récupérée. L'élution des protéines ayant une affinité avec la phase se fait avec 3 mL de tampon d'élution Tris.HCl 20 mM, NaCl 0.15 M, MnCl₂ 1 mM, CaCl₂ 1 mM, méthyl- α -D-pyranoside 200 mM à pH 7,4.

Enrichissement sur Concanavalline A au niveau peptidique

20 µg de protéines du surnageant de culture de *Mycobacterium tuberculosis* en milieu Sauton ont été déposés sur un gel SDS-Page, la migration a été stoppée lorsque les protéines

ont traversé la frontière entre le gel de concentration et le gel de séparation. Une bande a été découpée et le traitement biochimique décrit précédemment pour l'obtention du digestat peptidique a été réalisé.

30 µL de phase de concanavaline A greffée sur des billes d'agarose ont été prélevés. La phase a été lavée 3 fois consécutives avec 160 µL du tampon de lavage.

Le digestat peptidique est repris dans 400 µL de tampon de lavage et mis en présence de la lectine en suspension. Le mélange est incubé pendant 2H à 4°C. A l'issue de l'incubation, le mélange est centrifugé à 5000g pendant 15 minutes. Le surnageant est éliminé et la phase est récupérée. Cette phase est lavée 3 fois avec 60 µL de tampon de lavage, et les peptides retenus sont ensuite élués avec 40 µL de tampon d'élution.

Enrichissement sur phase HILIC au niveau peptidique

De la même façon que précédemment, 20 µg de protéines du surnageant de culture de *Mycobacterium tuberculosis* en milieu Sauton ont été digérés *in gel*.

Le digestat a ensuite été repris dans 20 µL d'une solution 80% ACN, 0,5% FA. Ce digestat est pipeté dans un cône d'une capacité de 10 µL contenant 30 µg de phase HILIC (Proteabio), préalablement équilibré avec de la solution 80% ACN, 0,5% FA. La phase est lavée 3 fois avec 20 µL de la solution 80% ACN, 0,5% FA. La fraction peptidique retenue est ensuite élue avec 2 µL d'une solution 99,5% H₂O, 0,5% FA avant d'être amenée à sec.

V. Analyse des protéines entières

Analyse nanoLC-MS

L'analyse nanoLC-MS des échantillons de protéines entières est également réalisée grâce à un système de nano-HPLC (U3000, Dionex) couplé à un spectromètre de masse (LTQ-Orbitrap Velos avec module ETD, Thermo Fischer Scientific). Les extraits protéiques secs sont repris dans 14 µL de solvant A'' (10% ACN, 0.05% acide trifluoro acétique (TFA)). La séparation des différentes espèces protéiques se fait sur une colonne analytique de phase inverse C4 (2,1 mm ID x 22 cm Aquapore, Brownlee) équilibrée avec 95 % de solvant A (5% ACN, 0,2 % FA) et 5% de solvant B (80% ACN, 0,2% FA). L'élution des protéines est réalisée par un gradient croissant de 5% à 50 % de solvant B sur 95 min à un débit de 300 nL/min. L'acquisition est également pilotée par le logiciel Xcalibur (Thermo Fischer Scientific). Les scans MS sont enregistrés dans l'analyseur Orbitrap avec une résolution de 100000 à m/z=400. Les spectres MS d'intérêt peuvent ensuite être manuellement sélectionnés pour être moyennés dans XCalibur. Chaque spectre moyenné ou non est ensuite déconvolué grâce au module Xtract implémenté dans XCalibur.

Analyse nanoLC-MS/MS, mode CID, HCD et ETD

Pour chaque analyse, le m/z d'un des états de charge d'une des protéoformes est entrée dans XCalibur et les ions possédant ce m/z vont être sélectionnés pour la fragmentation avec une fenêtre d'isolation de 10 m/z . Les spectres MS/MS sont acquis avec une résolution de 100000. Pour les analyses avec fragmentation ETD, un temps d'activation de 50 ms a été utilisé.

Le retraitement des données se fait grâce au logiciel ProSightPC (Thermo Scientific). Au sein de ce logiciel, chacun des spectres MS/MS est également déconvolué avec le module Xtract. Nous avons choisi le mode de retraitement « single protein mode » pour lequel la séquence de la protéine est à entrer manuellement. Les masses monoisotopiques des fragments théoriques sont alors calculées à partir de cette séquence. La liste des masses monoisotopiques théoriques est alors confrontée à la liste des masses monoisotopiques expérimentales, avec une tolérance choisie de 5 ppm. Les correspondances entre masses théoriques et masses expérimentales sont ensuite données.

PUBLICATION & COMMUNICATIONS

Liste des communications

Poster présenté au congrès 2012 de la Société française d'Electrophorèse et d'Analyse Protéomique à Rouen: Tonini L., A. Stella, C. Froment, C. Pichereaux, W. Malaga, M. Riviere and O. Schiltz. « Mass spectrometry characterization of an unusual glycosylation pattern of the *Mycobacterium smegmatis* O-mannosylated Fasciclin protein ».

Poster présenté à l'école d'été de protéomique de Brixen ainsi qu'au congrès de la Société française d'Electrophorèse et d'Analyse Protéomique à Saint-Malo en 2013:

Tonini L., M. Beau, CF. Liu, A. Stella, C. Froment, D. Bouyssie, W. Malaga, C. Guilhot, O. Schiltz and M. Riviere. « *Mycobacterium smegmatis* glycoproteomics analysis and full extent determination of the Fasciclin protein unusual glycosylation pattern ».

Bacterial protein-O-mannosylating enzyme is crucial for virulence of *Mycobacterium tuberculosis*

Chia-Fang Liu^{a,b,1}, Laure Tonini^{a,b,1}, Wladimir Malaga^{a,b}, Mathilde Beau^{a,b}, Alexandre Stella^{a,b}, David Bouyssie^{a,b}, Mary C. Jackson^c, Jérôme Nigou^{a,b}, Germain Puzo^{a,b}, Christophe Guilhot^{a,b}, Odile Burlet-Schiltz^{a,b}, and Michel Rivière^{a,b,2}

^aCentre National de la Recherche Scientifique, Institut de Pharmacologie et de Biologie Structurale, F-31077 Toulouse, France; ^bUniversité de Toulouse, Université Paul Sabatier, Institut de Pharmacologie et de Biologie Structurale, F-31077 Toulouse, France; ^cMycobacteria Research Laboratories, Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO 80523

Edited by Barry R. Bloom, Harvard School of Public Health, Boston, MA, and approved March 8, 2013 (received for review November 13, 2012)

A posttranslational protein O-mannosylation process resembling that found in fungi and animals has been reported in the major human pathogen *Mycobacterium tuberculosis* (Mtb) and related actinobacteria. However, the role and incidence of this process, which is essential in eukaryotes, have never been explored in Mtb. We thus analyzed the impact of interrupting O-mannosylation in the nonpathogenic saprophyte *Mycobacterium smegmatis* and in the human pathogen Mtb by inactivating the respective putative protein mannosyl transferase genes *Msmeg_5447* and *Rv1002c*. Loss of protein O-mannosylation in both mutant strains was unambiguously demonstrated by efficient mass spectrometry-based glycoproteomics analysis. Unexpectedly, although the *M. smegmatis* phenotype was unaffected by the lack of manno-proteins, the Mtb mutant had severely impacted growth in vitro and in cellulo associated with a strong attenuation of its pathogenicity in immunocompromised mice. These data are unique in providing evidence of the biological significance of protein O-mannosylation in mycobacteria and demonstrate the crucial contribution of this protein posttranslational modification to Mtb virulence in the host.

protein glycosylation | O-glycosylation | proteomics

With more than three deaths per minute, *Mycobacterium tuberculosis* (Mtb) is still a major public health threat, exacerbated by the emergence of multi- or extensively drug-resistant Mtb strains. This situation reflects the weakness of the therapeutic arsenal and highlights the urgent need for a comprehensive understanding of the determinants of Mtb pathogenicity for alternative therapeutics to fight tuberculosis.

We recently described Mtb manno-proteins as an emerging class of bacterial adhesins that contribute to Mtb infectiousness through binding to host innate-immune system receptors, including lung surfactant protein A (SP-A) (1) and DC-SIGN (2), which are considered to be the receptors preferentially used by mycobacteria to enter target cells and evade host defense mechanisms (3). Here, we considered the systemic impact of protein mannosylation on the survival and virulence of Mtb. Indeed the role of the protein glycan chains remains elusive (4), although it has been established that the immunodominant Mtb-secreted alanine and proline-rich antigen (Apa) interacts with the host's lectin receptors through its mannosyl appendages, which indirectly contribute to the colonization and invasion of the host cell. Moreover, changes in the mannosylation pattern of the *Mycobacterium bovis* bacillus Calmette–Guérin Apa alter its ability to stimulate CD4⁺ and CD8⁺ T lymphocyte responses (5, 6), contributing to the protective properties of bacillus Calmette–Guérin vaccination against tuberculosis (7). In addition, glycosylation (most probably O-mannosylation) of the Mtb lipoprotein LpqH modulates its antigenic processing (8), and glycan decorations of the *Mycobacterium leprae* lipoprotein LprG have been reported to be indispensable for MHC II-restricted T-cell activation in patients with lepromatous leprosy (9). However, there are still insufficient data to determine clearly the roles and overall

influence of mannosylation of mycobacterial proteins on Mtb physiology and infectiousness.

In eukaryotes, protein O-mannosylation is an essential process catalyzed by membrane-associated protein O-mannosyl transferases (PMTs), which transfer mannose residues from lipid carriers (dolichyl-phosphate) to proteins (10). An Mtb genome search for sequence homology with the eukaryotic PMTs identified *Rv1002c* as the unique putative PMT gene (11). Ectopic expression of *Rv1002c* in *Mycobacterium smegmatis* confirmed its mannosyl transferase function and demonstrated that this membrane-associated activity is coupled to the Sec-dependent protein export system, suggesting that O-mannosylation should affect only extracytoplasmic proteins (11). As for the eukaryotic PMTs, *Rv1002c* is categorized as a gene essential for optimal growth of *M. tuberculosis* in vitro (12, 13). This assumption is corroborated by the high conservation of *Rv1002c* homologs throughout the mycobacterial genus and, in particular, in the *M. leprae* genome, which is considered as the minimal set of genes essential for mycobacteria survival (14). On the other hand, PMT knockout mutants of the phylogenetically related *Streptomyces coelicolor* (15) or *Corynebacterium glutamicum* (16) are viable and suggest that this posttranslational modification is dispensable in the actinomycetaceae family. However, because of the proposed similarity of the protein O-mannosylation mechanism in eukaryotes and Mtb, we hypothesized that the conservation in pathogenic mycobacteria of such cell energy-consuming process should result from evolutionary selection of a protein modification required for the parasite to persist in its natural biotope: the infected host. To test this hypothesis we interrupted the putative PMT genes in the nonpathogenic saprophyte *M. smegmatis* and the human pathogen Mtb, and we verified whether posttranslational O-mannosylation of bacterial proteins is essential for Mtb. Using an efficient and original mass spectrometry-based glycoproteomic approach to tackle the highly challenging molecular analysis of the posttranslational glycosyl protein modifications, we confirmed the interruption of the mannosylation process and analyzed the impact of this interruption on in vitro and ex vivo phenotype and on virulence of Mtb in mice.

Results and Discussion

***M. smegmatis* *Msmeg_5447* Gene Coding for the Putative PMT Is Not Essential.** An in silico search for Mtb *Rv1002c* gene homologs (Fig. S1B) in the complete genome of the nonpathogenic saprophyte

Author contributions: M.R. designed research; C.-F.L., L.T., M.B., and M.R. performed research; W.M., A.S., D.B., M.C.J., J.N., and G.P. contributed new reagents/analytic tools; C.G., O.B.-S., and M.R. analyzed data; and M.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹C.-F.L. and L.T. contributed equally to this work.

²To whom correspondence should be addressed. E-mail: michel.riviere@ipbs.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1219704110/-DCSupplemental.

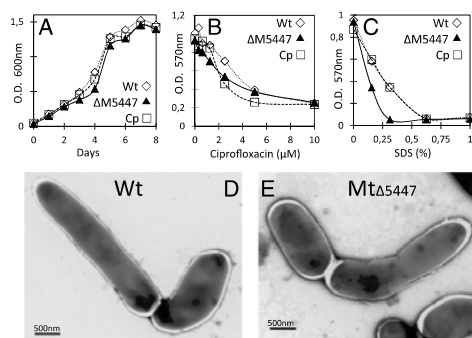


Fig. 1. The *Msmeg_5447* gene encoding the putative ortholog of the Mtb *Rv1002c* mannosyl transferase gene is dispensable for *M. smegmatis* growth in vitro. Effects of *Msmeg_5447* inactivation on the growth (A) and the sensitivity of the *M. smegmatis* ΔM5447 mutant to the antituberculosis drug ciprofloxacin (B), and detergent (SDS)-induced cell wall stress (C). Growth and survival were followed by monitoring the OD at 600 nm (A) or the level of MTT reduction by the metabolically active cells at 570 nm (B and C). (D and E) Transmission electron micrographs of negatively stained *M. smegmatis* wild-type (Wt) and ΔM5447 mutant cells.

M. smegmatis mc²-155 revealed a unique ORF (*Msmeg_5447*: 1,551 pb) coding for a putative 516-amino acid membrane protein sharing 72% identity and 81% homology over 489 amino acids with the Mtb PMT sequence (Fig. S1A). An *M. smegmatis* insertion mutant, ΔM5447, was constructed by disruption of *Msmeg_5447* by allelic exchange (Fig. S1C and D). Phenotypic analysis of the mutant clearly indicated the dispensability of the putative PMT function for *M. smegmatis*, which contrasts with the essential nature of PMT in eukaryotes. Indeed, there was no significant difference in growth between the WT and the insertion mutant (ΔM5447) (Fig. 1A). Similarly, the relative susceptibility of both strains to fluoroquinolone antibiotic (ciprofloxacin) (Fig. 1B) was similar, suggesting that *Msmeg_5447* inactivation had little effect on metabolism or cell wall permeability to drugs. In contrast, tolerance to cell wall stress induced by the SDS detergent was slightly reduced in the ΔM5447 mutant in comparison with the WT (Fig. 1C), which was fully reversed in the complemented mutant strain. The increase in sensitivity to SDS may reflect changes in cell wall organization or composition of the ΔM5447 mutant, weakening its physical and chemical protective properties; however, transmission electron microscopy did not reveal any noticeable differences in the cell wall aspect or cell shape (Fig. 1D and E) of the WT and the ΔM5447 mutant.

Evidence for the Putative PMT Function of the *Msmeg_5447* Gene Product. In the absence of major phenotypic changes associated with inactivation of *Msmeg_5447*, we analyzed the mutant and WT secreted proteomes to search for potential alterations in protein glycosylation. Comparative SDS/PAGE analysis (Fig. 2A) of the culture filtrates revealed a major difference corresponding to the downsizing of an intense protein gel band from 30 to 25 kDa, consistent with a glycosylation defect in the mutant. Complementation of the mutant strain with *Msmeg_5447* led to restoration of the original mobility of the gel band at around 30 kDa, thus confirming that this alteration is related to the *Msmeg_5447* gene-product deficiency. Proteomic analysis by nanoLC-MS/MS of the in-gel trypsin digest of the corresponding excised gel bands allowed the identification of several peptides attributable to the predicted *M. smegmatis* fascilin domain protein

(FasC) (Fig. 2B and Table S1). A recombinant C-terminal histidine tagged FasC (rFasC) was then constructed for expression in *M. smegmatis*, followed by further purification for precise structural characterization. Carbohydrate content capillary zone electrophoresis analysis (Fig. 2C) and α-exo-mannosidase susceptibility study (Fig. 2D) confirmed that the purified rFasC is glycosylated by α-mannose residues exclusively. Mass spectrometry analysis of the rFasC trypsin digest afforded 42% amino acid sequence coverage and determined the location of the glycosylation (up to four hexose residues) on the N-terminal P_{30–55}

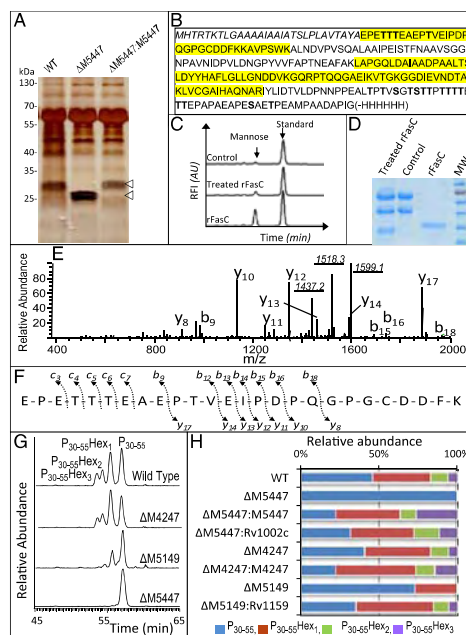


Fig. 2. Inactivation of the *Msmeg_5447* gene interrupts the O-mannosylation of the *M. smegmatis* FasC protein. (A) Identification of the *Msmeg_5447* protein target by silver nitrate-stained SDS/PAGE of the culture filtrate protein extract from *M. smegmatis* WT, ΔM5447 mutant, and complemented mutant (ΔM5447:M5447) revealing the electrophoretic mobility alteration of a major protein gel band (arrows). (B) Amino acid sequence of the recombinant His-Tagged rFasC (SignalP 3.0 predicted signal peptide is in italics; NetOGlyc 3.1 predicted S and T glycosylation sites are boldface; Histidine tag is into brackets; yellow highlights the amino acid sequence covered by mass spectrometry analysis). (C) Carbohydrate content analysis and (D) Coomassie blue-stained SDS/PAGE of the rFasC, the α-exo-mannosidase-treated rFasC, and the α-exo-mannosidase sham control (Standard: manno-heptose). (E) Identification of the triglycosylated rFasC N-terminal peptide P_{30–55} by CID MS/MS [(P_{30–55}Hex₃+3H)³⁺ at *m/z* 1680.23]. Underlined masses correspond to fragment ions resulting from one to three neutral losses of hexose from the parent ion. (F) FasC N-terminal peptide P_{30–55} sequence reporting the major informative peptide cleavages (b, c, and y peptide fragment ions) observed in the positive CID and ETD MS/MS spectra (Fig. S2). (G) Reconstructed ion chromatograms (mass tolerance: 10 ppm, Sum of [M+3H]³⁺ [M+2H]²⁺) of the native FasC P_{30–55} glycoforms detected in the culture filtrates of the WT and the mannosyl transferase mutants. (H) Cumulated normalized abundances of the native FasC P_{30–55} glycoforms observed in the mannosyl transferase mutants and their respective complemented strains.

peptide (Fig. 2 E and F, and Tables S1 and S2). Accurate identification of the glycosylated threonine residues was achieved by complementary electron transfer dissociation (ETD) MS/MS analysis (17). Interestingly, we found, that in the rapid growing saprophyte *M. smegmatis*, as reported in other mycobacteria, the carbohydrate motifs are not randomly distributed along the four potential hydroxylated amino acid acceptors (Thr₃₃, Thr₃₄, Thr₃₅, Thr₄₀) of the P₃₀₋₅₅ peptide but are exclusively located on adjacent Thr residues (Thr₃₃-Thr₃₄ doublet) (Fig. S2 and Table S3) (18–20). The identification and the determination of the glycosylation sites of FasC, which is the first ever mannoprotein characterized in *M. smegmatis*, allowed us to explore the alteration of protein O-mannosylation in the mutant by monitoring specifically the glycosylation status of the FasC P₃₀₋₅₅ peptide by mass spectrometry. Searching for the molecular ions of the FasC P₃₀₋₅₅ reporter peptide in the ΔM5447 mutant culture filtrate by nanoLC-MS analysis readily revealed the absence of peptide glycosylation (Fig. 2G, Bottom trace), in agreement with the apparent mass reduction of the mutant FasC protein observed by SDS/PAGE (Fig. 2A). Interestingly, the glycosylation defect was reversed by complementation with the *Msmeg_5447* gene as well as with the Mtb *Rv1002c* ortholog (Fig. 2H), providing evidence of functional identity of these genes.

Initiation and Elongation of the FasC Mannosyl Appendages Requires *Msmeg_5447* PMT and PimE. O-glycosylation in bacteria has been described to proceed either by *en bloc* transfer (of a lipid-linked oligosaccharide to the protein) or by a stepwise processing initiated by attachment of an initial glycosyl residue to the target serine or threonine and continued by elongation of the oligosaccharide chain by successive glycosylation. By analogy with the eukaryote PMT-mediated process, protein O-mannosylation in Mtb has been considered to proceed according to the latter mechanism (11), which is the exclusive rule for O-glycosylation in eukaryotes. However, the mycobacterial enzymes involved in the elongation process are not known. The oligomannosyl chains of the Mtb proteins have been found elongated through α1–2 interglycosidic linkages (18, 20), as in the polar phosphatidyl inositol

mannosides (PIM6) or the mannose branches of the lipomannan and lipoarabinomannan core (21, 22). Thus, mutants of the α1–2 mannosyl transferases *Msmeg_5149* (pimE) (23) and *Msmeg_4247*, respectively responsible for the elongation of PIMs and lipoglycans in *M. smegmatis*, were constructed to verify the possible involvement of these enzymes in the elongation of the FasC mannosyl chain. As revealed by the reporter FasC P₃₀₋₅₅ glycoforms detected by nanoLC-MS in the culture filtrate of the ΔM4247 mutant (Fig. 2G), disruption of *Msmeg_4247* failed to induce any significant modification of the distribution of the reporter glycopeptide, suggesting that this α1–2 mannosyl transferase does not contribute to the elongation of FasC mannosyl chains. In contrast, inactivation of the PimE mannosyl transferase resulted as expected in the loss of the polar PIMs (Fig. S3B), but also in the total extinction of the triglycosylated form of the FasC P₃₀₋₅₅ peptide with the presence of trace amounts (<1.5%) of the diglycosylated isoform (Fig. 2G). This glycosylation defect was reversed by complementation with the Mtb *pimE* (*Rv1159*) gene (Fig. 2H), which confirms the involvement of the *Msmeg_5149* mannosyl transferase in the elongation of FasC P₃₀₋₅₅ oligomannosyl appendages and the functional identity between *Msmeg_5149* and *Rv1159*. These results show that, protein mannosylation in *M. smegmatis* occurs through a stepwise process initiated by the transfer of the primary mannose residue onto protein by the PMT encoded by *Msmeg_5447* followed by the elongation of the mannosyl chain through α1–2 linkages by PimE. It is noteworthy that the pleiotropic function of the *M. smegmatis* PimE evokes the shared glycosylation machinery for glycolipids and glycoproteins reported in the Gram-negative pathogens (24). However, unlike the Gram-negative bacteria oligosaccharidyl-transferase, which catalyzes the oligosaccharide transfer *en bloc* from glycolipids to acceptor glycoproteins, PimE catalyzes the stepwise oligomannosyl chain elongation indifferently on manno-proteins and manno-lipids, proving a remarkable relaxed specificity. In contrast, the PMT is protein-specific and unlikely to contribute to the PIM biosynthesis (Fig. S3C).

Table 1. *M. smegmatis* glycoproteins identified by glycopeptidomics analysis of the WT strain culture filtrate

Protein description	Access no.*	Gene no.	MM†	Peptide sequence	Hexose no.
Fascidin domain protein	A0R2Q4	5196	3020.3	(A)EPETTTEAEPTVEIPDPQGGPGC [‡] DDFK(K)	1–3 Hex
			3148.4	(A)EPETTTEAEPTVEIPDPQGGPGC [‡] DDFK(A)	1–3 Hex
ABC-type amino acid transport system secreted component	A0QXB0	3235	1867.8	(S)GEGGGESSPESTQAASGAK(V)	1–2 Hex
			1954.8	(A)SGEGGGESSPESTQAASGAK(V)	1–2 Hex
			2025.9	(C)ASGEGGGESSPESTQAASGAK(V)	1–3 Hex
Putative protein	A0R6M4	6601	2574.3	(K)STTPPTPLPPPLPAEVGGSAK(A)	3–5 Hex
Putative uncharacterized protein	A0QYI9	3674	2777.3	(K)VILMLDC [§] TEAAQQAQDTAVSGGPR(V)	1 Hex
			3068.5	(R)VGSN [¶] GVTTVTPTPVPIAGAPGAGTPPPA(-)	3–4 Hex
			1876.9	(R)AQPVTPTQSDTVPGR(Y)	2 Hex
Putative uncharacterized protein	A0QRP2	1188	3702.9	(R)YLPVPVPHDPVALPVTLTTPPAQPVAPNTR(T)	2 Hex
			1904.0	(T)TTVPAPAPTTAPVSHGPAA(-)	1 Hex
Immunogenic protein MPT63	A0R3B5	5412	2389.2	(K)DLLVWDKPAASATAPSGTQSR(P)	1 Hex
			4227.9	(R)ATDLPAATAAEATEATATDAETTPVTEETPAAPAEPTTPV(S)	1 Hex
			1994.0	(K)LTLLTTTATVNDEAQK(Q)	1 Hex
Putative protein	A0QSU1	1600	1492.7	(R)PAATSGSGTSASTPAR(T)	1 Hex
Immunogenic protein MPT63	A0QQP4	828	3127.5	(K)WVPVIETDPKPFDC [‡] NDIPDIVIER(I)	1 Hex
Putative protein	A0QVP4	2645	2119.0	(R)PGVGVPVPVTDAPPEMMPPA(-)	1 Hex
Alanine and proline-rich secreted protein apa	A0QYD3	3618			

‡Boldface T and S correspond to potential glycosylation sites according NetOGlyc (v3.1) predictions.

*UniProtKB accession number.

†Experimental mass of the glycosylated peptide of lower glycosylation degree.

‡Carboxymethylated cysteine.

§Cysteine acrylamide adduct.

¶Deamidation.

mutant strain (Fig. 3D). Further normalization of the peptide glycoform abundances (relative to the closely eluted prototypic P_{174–184} Apa peptide used as internal standard) supported this observation by showing that the mannosylation defect was correlated to a relative increase in the abundances of the unmodified peptides (Fig. 3C, Left). These results clearly demonstrate that disruption of *Rv1002c* in Mtb abolishes protein mannosylation, similar to the situation in the model strain *M. smegmatis*.

Maintenance of the $\Delta Rv1002c$ Mutant in Macrophages Is Compromised.

We next investigated the impact of the deficiency in *Rv1002c*-dependent protein *O*-mannosylation on the interaction of Mtb with the host. The growth defect on solid medium induced by the inactivation of *Rv1002c* clearly impaired the use of the conventional counting of colony forming units to monitor the intracellular survival of the mutant. Fluorescent Mtb strains were thus generated by expressing the *gfp* gene from a replicative plasmid in the parental WT, the $\Delta Rv1002c$ mutant, and the complemented mutant to evaluate the consequences of the inactivation of the *Rv1002c* gene on the ability of the mutant to infect and proliferate in host cells. Resting cultures of murine alveolar macrophage cell line and primary human monocyte-derived macrophages were infected for 1 h at a multiplicity of infection (MOI) of one bacterium per macrophage. In the murine macrophages, the initial rate of infection is slightly lower for the mutant than for the WT or

complemented strains (Fig. 4A). This finding is consistent with the view that surface-exposed glycoproteins such as Apa may act as bacterial adhesins for the phagocyte membrane C-type lectin receptors (1). However, the absence of similar statistical difference with the human macrophages (Fig. 4C) not only reveals subtle differences in the impact of the surface manno-proteins on the initial uptake by mouse or human phagocytes, but also suggests that the overall divergence observed between the mutant and the WT intracellular behavior is not formally related to an altered initial capacity of the mutant to interact with the host cell. Indeed, the main difference between the $\Delta Rv1002c$ mutant and the Mtb WT or complemented mutant strains was the almost total incapacity of the mutant to proliferate within the phagocytes. The Mtb WT strain proliferated in either macrophage cell culture, as shown by both the number of infected macrophages and the bacterial load values, which increased perceptibly as early as 48 h postinfection and had almost doubled at day 7 (Fig. 4). On the other hand, the number of $\Delta Rv1002c$ mutant fluorescent cells per infected macrophage remained almost stable, and the number of macrophages infected by the mutant cells declined by about 50% over the 7-d observation period. These findings indicate a possible block in the capacity of the mutant to multiply within the macrophage. This phenotype was nearly completely reversed in the complemented strain, confirming that the functional disruption of the putative PMT encoded by *Rv1002c* dramatically compromises the mutant ability to survive and multiply in phagocytes.

Virulence of the $\Delta Rv1002c$ Mutant Is Attenuated in SCID Mice. Given the incapacity of the mutant to proliferate in macrophages, we predicted a severe loss of pathogenicity in vivo, even in the absence of host-acquired immunity. To test this theory, virulence assays were carried out by infecting immunocompromised SCID mice via the intranasal route with 10^3 cells of the Mtb parental WT, the $\Delta Rv1002c$ mutant, or the complemented mutant. Control mice infected with the virulent WT Mtb strain did not survive for more than 22 d, whereas no death was observed until day 47 (median survival time = 63 d, $n = 9$) in the mice infected with the mutant strain (Fig. 4E). By increasing the dose of $\Delta Rv1002c$ mutant, it was clear that survival time was dependent on the initial bacterial load used to infect the mice (Fig. 4F). A rough estimate of the relative pathogenicity of the mutant, obtained by extrapolating the median survival time as a function of the initial bacterial load, indicated a virulence attenuation of about 10^{-3} to 10^{-4} [10^3 to 10^5 -fold more mutant bacteria are required to give a median survival time (~ 22 d) equivalent to that observed with 10^3 WT Mtb cells]. Interestingly, complementation of the mutant with the *Rv1002c* gene readily restored mutant pathogenicity, as demonstrated by the death of all of the mice infected with the complemented strain by day 22 (Fig. 4E). This finding clearly confirms that the attenuation of the mutant virulence is related to the inactivation of *Rv1002c* gene coding for the Mtb PMT.

Conclusion and Future Prospects. Taken together, our findings reveal that although mycobacterial protein *O*-mannosylation occurs by a similar mechanism to that in eukaryotes (11, 29), it strongly differs in its biological impact. Unlike in eukaryotes, this process is dispensable, at least in laboratory culture conditions, for nonpathogenic saprophyte and pathogenic parasite mycobacteria. On the other hand, we demonstrate in vivo that interruption of protein *O*-mannosylation crucially attenuates Mtb virulence in an immunocompromised mouse model. Therefore, we speculate that decorations of the Mtb surface and secreted proteins with $\alpha 1-2$ linked mannobioses mimicking the high mannose branch motifs of the eukaryotic intracellular glycoproteins are not fortuitous and may be crucial for Mtb to establish effectively parasitic relationships with its eukaryotic hosts. Further explorations need to be undertaken to determine the molecular basis

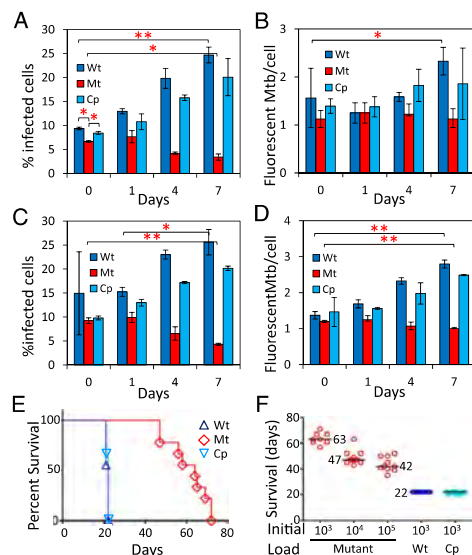


Fig. 4. *Rv1002c*-dependent protein *O*-mannosylation is required for Mtb intracellular persistence and proliferation and for full virulence in SCID mice. (A–C) Persistence and (B–D) intracellular proliferation of GFP-tagged Mtb wild-type (Wt), $\Delta Rv1002c$ mutant (Mt), and complemented mutant (Cp) in the mouse alveolar macrophage cell line (A and B) and primary human blood monocyte-derived macrophages (C and D). [Data (Table S5) were analyzed by one-way Anova tests: * $P < 0.05$, ** $P < 0.01$]. Survival of immunocompromised SCID mice infected intranasally with 10^3 Mtb WT, $\Delta Rv1002c$ mutant, or complemented mutant cells (E), or with 10^3 , 10^4 , or 10^5 $\Delta Rv1002c$ mutant cells or 10^3 Mtb WT and complemented mutant cells (F). Numbers correspond to the median survival time.

of the phenotypic alterations of the $\Delta Rv1002c$ mutant and to decipher the relationships between the interruption of *Rv1002c*-dependent protein O-mannosylation, the colony-forming deficiency, and the virulence attenuation. Our findings are unique in providing evidence for the crucial influence of mycobacterial protein O-mannosylation on Mtb physiology and virulence, and open the door for the exhaustive exploration of the PMT target proteome and analysis of the roles of mannosylated proteins in tuberculosis pathophysiology. Such holistic consideration of protein O-mannosylation-dependent Mtb persistence and proliferation should provide new perspectives in the search for original strategies to fight tuberculosis.

Materials and Methods

The *M. smegmatis* *Msmeg_5447* and *M. tuberculosis* *Rv1002c* mutants (Figs. S1 and S4, and Table S4) were complemented with pMV361 derivative carrying an *Hyg^r* cassette and the *Msmeg_5447* or *Rv1002c* genes under the control of the *pBlafA** promoter. The *FasC Msmeg_5196* gene was fused to a polyhistidine tag within *Hyg^r-pMIP12* derivative under the control of *pBlafA** promoter. *M. smegmatis* mc²155 and *M. tuberculosis* H37Rv cells were grown for 5–7 d and 20–30 d ($OD_{600} = 0.5$ – 0.6) (30). Mtb culture mediums were doubly filtered on 0.22- μ m membrane and concentrated on Vivaspin 5 kDa, rFasC was purified by FPLC before carbohydrate analysis and α -exomannosidase treatment (1). NanoLC-MS/MS analyses were performed on an Ultimate3000 system (Dionex) coupled to an LTQ-Orbitrap XL or an ETD-enabled LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific), respectively, for CID or ETD experiments. Data analysis and database searches were performed according to

journal guidelines using Mascot Daemon against actinobacteria entries in Sprot-Trembl_20090707 database and the in-house developed software MFPaQ v4.0.0 (31, 32). The detection of peak signatures corresponding to carbohydrate neutral loss in CID MS/MS spectra has been performed using an in-house developed Perl script. Human peripheral blood mononuclear cells (hPBMC) were prepared from nontuberculous control donor blood (33). hPBMC and murine macrophages (34) were infected by GFP-mycobacteria at a MOI = 1 for 1 h at 37 °C. Cells collected at 0, 24, 96 h, and 7 d were fixed with paraformaldehyde and green fluorescent intracellular mycobacteria were quantified by fluorescence microscopy (Table S5). Animal studies were approved by the Comité d'Ethique Midi-Pyrénées and conducted according to the Centre National de la Recherche Scientifique housing and care guidelines for laboratory animals. Groups of nine female SCID mice were infected intranasally with 10^3 cfu (otherwise stated) of either *M. tuberculosis* H37Rv strain and monitored daily. Animals showing signs of illness were killed.

ACKNOWLEDGMENTS. We thank Dr. P. Constant (MTT micro-assay), F. Moreau and L. Lepourry (animal experiments), C. Froment (electron transfer dissociation mass spectrometry), Dr. M. Gilleron (phosphatidyl inositol mannoside analysis), and Dr. G. Larrrouy-Maumus (Institut de Pharmacologie et de Biologie Structurale, Toulouse, France), the microscopy facilities IFR109 (Institut de Biologie Cellulaire et de Génétique, Toulouse, France), and Dr. D. Kaur ($\Delta M4247$ and $\Delta M5149$ mutants) (Colorado State University). This work was supported by the Agence Nationale de la Recherche Grant 11BSV6016 (to M.R.); the National Institute of Allergy and Infectious Diseases/National Institutes of Health Grant A1064798 (to M.C.J.); the French Ministry of Research (Investissements d'avenir, Proteomics French Infrastructure) and the Fonds Européens de Développement Régional and the Région Midi-Pyrénées (O.B.-S.); the Centre National de la Recherche Scientifique (M.B.); the Région Midi-Pyrénées (L.T.); and the National Science Council, Taiwan NSC98-2311-B-002-MY2-001-MY2 (to C.-F.L.).

- Ragas A, Roussel L, Puzo G, Rivière M (2007) The *Mycobacterium tuberculosis* cell-surface glycoprotein apa as a potential adhesin to colonize target cells via the innate immune system pulmonary C-type lectin surfactant protein A. *J Biol Chem* 282(8): 5133–5142.
- Pitarque S, et al. (2005) Deciphering the molecular bases of *Mycobacterium tuberculosis* binding to the lectin DC-SIGN reveals an underestimated complexity. *Biochem J* 392(Pt 3):615–624.
- Nguyen L, Pieters J (2005) The Trojan horse: Survival tactics of pathogenic mycobacteria in macrophages. *Trends Cell Biol* 15(5):269–276.
- Espitia C, Servin-González L, Mancilla R (2010) New insights into protein O-mannosylation in actinomycetes. *Mol Biosyst* 6(5):775–781.
- Romain F, et al. (1999) Deglycosylation of the 45/47-kilodalton antigen complex of *Mycobacterium tuberculosis* decreases its capacity to elicit in vivo or in vitro cellular immune responses. *Infect Immun* 67(11):5567–5572.
- Horn C, et al. (1999) Decreased capacity of recombinant 45/47-kDa molecules (Apa) of *Mycobacterium tuberculosis* to stimulate T lymphocyte responses related to changes in their mannosylation pattern. *J Biol Chem* 274(45):32023–32030.
- Kumar P, Amara RR, Challa VK, Chadda VK, Satchidanandam V (2003) The Apa protein of *Mycobacterium tuberculosis* stimulates gamma interferon-secreting CD4+ and CD8+ T cells from purified protein derivative-positive individuals and affords protection in a guinea pig model. *Infect Immun* 71(4):1929–1937.
- Herrmann JL, O'Gaora P, Gallagher A, Thole JE, Young DB (1996) Bacterial glycoproteins: A link between glycosylation and proteolytic cleavage of a 19 kDa antigen from *Mycobacterium tuberculosis*. *EMBO J* 15(14):3547–3554.
- Sieling PA, et al. (2008) Conserved mycobacterial lipoglycoproteins activate TLR2 but also require glycosylation for MHC class II-restricted T cell activation. *J Immunol* 180(9):5833–5842.
- Strahl-Bokinger S, Gentzsch M, Tanner W (1999) Protein O-mannosylation. *Biochim Biophys Acta* 1426(2):297–307.
- VanderVen BG, Harder JD, Crick DC, Belisle JT (2005) Export-mediated assembly of mycobacterial glycoproteins parallels eukaryotic pathways. *Science* 309(5736): 941–943.
- Griffin JE, et al. (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog* 7(9):e1002251.
- Sassetti CM, Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci USA* 100(22):12989–12994.
- Vissa VD, Brennan PJ (2001) The genome of *Mycobacterium leprae*: A minimal mycobacterial gene set. *Genome Biol* 2(8):S1023.
- Wehmeier S, et al. (2009) Glycosylation of the phosphate binding protein, PstS, in *Streptomyces coelicolor* by a pathway that resembles protein O-mannosylation in eukaryotes. *Mol Microbiol* 71(2):421–433.
- Mahne M, Tauch A, Pühler A, Kalinowski J (2006) The *Corynebacterium glutamicum* gene pmt encoding a glycosyltransferase related to eukaryotic protein-O-mannosyltransferases is essential for glycosylation of the resuscitation promoting factor (Rpf2) and other secreted proteins. *FEMS Microbiol Lett* 259(2):226–233.
- Zubarev RA, et al. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem* 72(3):563–573.
- Dobos KM, Khoo KH, Swiderek KM, Brennan PJ, Belisle JT (1996) Definition of the full extent of glycosylation of the 45-kilodalton glycoprotein of *Mycobacterium tuberculosis*. *J Bacteriol* 178(9):2498–2506.
- Michell SL, et al. (2003) The MPB83 antigen from *Mycobacterium bovis* contains O-linked mannose and (1–3)-mannobiose moieties. *J Biol Chem* 278(18):16423–16432.
- Sartain MJ, Belisle JT (2009) N-Terminal clustering of the O-glycosylation sites in the *Mycobacterium tuberculosis* lipoprotein SodC. *Glycobiology* 19(1):38–51.
- Chatterjee D, Lowell K, Rivoire B, McNeil MR, Brennan PJ (1992) Lipoarabinomannan of *Mycobacterium tuberculosis*, Capping with mannose residues in some strains. *J Biol Chem* 267(9):6234–6239.
- Hunter SW, Brennan PJ (1990) Evidence for the presence of a phosphatidylinositol anchor on the lipoarabinomannan and lipomannan of *Mycobacterium tuberculosis*. *J Biol Chem* 265(16):9272–9279.
- Morita YS, et al. (2006) PimE is a polyphosphatide-mannose-dependent mannosyltransferase that transfers the fifth mannose of phosphatidylinositol mannoside in mycobacteria. *J Biol Chem* 281(35):25143–25155.
- Nothhaft H, Szymanski CM (2010) Protein glycosylation in bacteria: Sweeter than ever. *Nat Rev Microbiol* 8(11):765–778.
- Zaia J (2010) Mass spectrometry and glycomics. *OMICS* 14(4):401–418.
- Banael N, et al. (2009) Lipoprotein processing is essential for resistance of *Mycobacterium tuberculosis* to malachite green. *Antimicrob Agents Chemother* 53(9): 3799–3802.
- Espitia C, Mancilla R (1989) Identification, isolation and partial characterization of *Mycobacterium tuberculosis* glycoprotein antigens. *Clin Exp Immunol* 77(3): 378–383.
- Dobos KM, Swiderek K, Khoo KH, Brennan PJ, Belisle JT (1995) Evidence for glycosylation sites on the 45-kilodalton glycoprotein of *Mycobacterium tuberculosis*. *Infect Immun* 63(8):2846–2853.
- Lommel M, Strahl S (2009) Protein O-mannosylation: Conserved from bacteria to humans. *Glycobiology* 19(8):816–828.
- Le Dantec C, Winter N, Gicquel B, Vincent V, Picardeau M (2001) Genomic sequence and transcriptional analysis of a 23-kilobase mycobacterial linear plasmid: Evidence for horizontal transfer and identification of plasmid maintenance systems. *J Bacteriol* 183(7):2157–2164.
- Mouton-Barbosa E, et al. (2010) In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. *Mol Cell Proteomics* 9(5):1006–1021.
- Bouyssié D, et al. (2007) Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: Application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol Cell Proteomics* 6(9): 1621–1637.
- Astarie-Dequeker C, et al. (1999) The mannose receptor mediates uptake of pathogenic and nonpathogenic mycobacteria and bypasses bactericidal responses in human macrophages. *Infect Immun* 67(2):469–477.
- Liu CF, Rivere M, Huang HJ, Puzo G, Wang JY (2010) Surfactant protein D inhibits mite-induced alveolar macrophage and dendritic cell activations through TLR signalling and DC-SIGN expression. *Clin Exp Allergy* 40(1):111–122.

Supporting Information

Liu et al. 10.1073/pnas.1219704110

SI Materials and Methods

Construction of *Mycobacterium smegmatis* and *Mycobacterium tuberculosis* Mutants. As shown in Table S4, the *Msmeg_5447* mutant was constructed using the method described by Pelicic et al. (1). Briefly a 2.1-kb fragment covering the *Msmeg_5447* gene was amplified by PCR from genomic DNA using primer 5447A and 5447B and cloned into pGEM-T (Promega) to give pWM150. A kanamycin resistance cassette formed by the Ω km cassette flanked by two *res* sites from transposon $\gamma\delta$ (2) was inserted at PmeI restriction site located within *Msmeg_5447* (Fig. S1C). An EcoRV restriction fragment containing *Msmeg_5447* disrupted by the *res*- Ω km-*res* cassette was recovered from this plasmid and inserted at the SmaI restriction site of the counter-selectable vector pJQ200 to yield pWM152. This plasmid was transferred by electroporation into *M. smegmatis* mc²155 (3) and allelic exchange mutants were selected on LB agar plates containing *K_m* and 4% (wt/vol) sucrose. *K_m*^R and *Suc^R* clones were analyzed by PCR using oligonucleotides 5447C, 5447D, *res2*, and *res1* (Fig. S1D). A clone exhibiting the expected PCR profile for an allelic exchange mutant was retained for further analysis and named PMM145.

The *Ts/sacB* method was used to achieve allelic replacement at the *Msmeg_5149* locus of *M. smegmatis* mc²155 (1). *Msmeg_5149* and flanking regions were PCR-amplified and a disrupted allele was obtained by replacing 750 bp of the coding sequence of this gene bracketed between two NruI sites by the same *K_m* resistance cassette. A similar approach was used to disrupt the *Msmeg_4247* gene.

The *M. tuberculosis* H37Rv (4) *Rv1002c* mutant was constructed using the strategy described by Bardarov et al. (5). A DNA fragment containing the *Rv1002c* gene was amplified by PCR from genomic DNA using oligonucleotides 1002C and 1002D and inserted within the vector pSC-B (Stratagene) to give plasmid pWM167. An internal *Rv1002c* fragment was removed by SphI-AatII digestion and substituted by the *res*- Ω km-*res* cassette to yield plasmid pWM169 (Fig. S4A). The NheI-XhoI fragment from pWM169 was then cloned within the cosmid vector pYUB854 (5). The resulting cosmid was then cut with PacI and ligated with the mycobacteriophage ϕ AE87 to give ϕ WM11. This phagemid was transferred by electroporation in *M. smegmatis* and phage particles were prepared as described previously (5). These particles were then used to infect *M. tuberculosis* H37Rv and *M. tuberculosis* allelic exchange mutants were selected on 7H11 agar plates supplemented with OADC and *K_m*. Several clones were analyzed using primers, *res1*, *res2*, 1002E, 1002F, 1002G, and 1002H. One clone, named PMM152, gave the pattern corresponding to allelic exchange and was retained for further analysis (Fig. S4B).

The *res*- Ω km-*res* cassette was recovered by transferring into PMM152 the plasmid pWM19 containing the resolvase gene of transposon $\gamma\delta$. One clone *K_m*^R, named PMM176, displaying a PCR profile consistent with excision of the *res*- Ω km-*res* cassette, was retained for further analysis.

Complementation of *M. smegmatis* or *M. tuberculosis* Mutant. The *Msmeg_5447* and *Rv1002c* gene were amplified by PCR from genomic DNA using primer 5447E/5447F and 1002A/1002B respectively. These ~1.5-kb fragments were sequenced and inserted into a pMV361 (6) derivative carrying a hygromycin cassette, *hyg*, and the pBlaF* promoter instead of the original *phsp60* promoter (7). The two resulting plasmids were named pWM159H (*Msmeg_5447*) and pWM158H (*Rv1002c*) respectively and

transferred by electroporation into *M. smegmatis* Δ M5149 or *M. tuberculosis* Δ Rv1002c mutants.

For complementation of the *M. smegmatis* Δ M5149 mutant, two plasmids were produced. First, the entire coding sequence of *Rv1159* was PCR-amplified and cloned into the expression vector pVV16 (8), yielding pVVRv1159. Second, *Msmeg_5149* gene was obtained by PCR amplification from *M. smegmatis* chromosomal DNA and inserted within the pMV361 derivative under the control of the pBlaF* promoter to form pWM292. These constructs were transferred into the *M. smegmatis* Δ M5149 mutant by electroporation.

Expression of FasC Fused with a Polyhistidine Tag. The *Msmeg_5196* gene encoding the fasciclin domain protein (FasC) was amplified from genomic DNA using primers 5196_A and 5196_B. The DNA fragment was sequenced and fused to a polyhistidine-encoding tag before insertion within a pMIP12 (9) derivative carrying a hygromycin cassette to give pWM218. In this construct, the expression of *Msmeg_5196* gene is under the control of the pBlaF* promoter.

Growth Conditions. *M. smegmatis* cells were grown aerobically to log-phase for 5–7 d at 37 °C in Middlebrook 7H9 broth supplemented with 0.2% glycerol and 0.05% tween 80 and *M. tuberculosis* strains were grown for 20–30 d (*OD*₆₀₀ = 0.5–0.6) on the same medium plus 2.5% (vol/vol) ADC (Albumin-Dextrose-Catalase, Becton Dickinson Microbiology System) enrichment. Search for native albumin free culture conditions for the *M. tuberculosis* strains were performed in 7H9 broth supplemented with dextrose (1 g/L), catalase (1.5 mg/L), albumin (2.5 g/L) or albumin trypsin digest [eq. 2.5 g/L albumin treated 4 h at 37 °C with 1% (wt/wt) trypsin in 50 mM NH₄HCO₃, pH 7.8]. Solid support culture were performed on Middlebrook 7H11 medium or Malachite green free Middlebrook 7H9 agar supplemented with 0.5% glycerol and 10% (vol/vol) OADC (Becton Dickinson Microbiology System). Hygromycin and Kanamycin were added at 50 mg/L and 40 mg/L for selection of the appropriate strains.

Drug and SDS Sensitivity Assays. These assays were realized in 96-well microtiter plates using the tetrazolium salt MTT colorimetric microassay based on the reduction by metabolically active cells of the [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide; Sigma] to formazan, as previously described (10). Briefly, 100 μ L of bacterial suspension (*OD* = 0.04) in 7H9 broth were incubated in presence of the appropriate SDS or drug concentration for 6 d before addition of 50 μ L of MTT (1 mg/mL) and measurement of the absorbance at 570 nm.

Transmission Electron Microscopy. For transmission electron microscopy, 5- to 7-d culture mycobacteria (*10*⁶ to *10*⁸) were washed three times in distilled water (17,000 \times g, 2 min) and suspended in a final volume of 50 μ L. For negative staining, 10 μ L of washed mycobacteria preparations were applied to collodion-carbon-coated copper grids (200 meshes). After 10 s, solution in excess was blotted off with filter paper and samples were stained with 1% (wt/vol) uranyl acetate for 30 s. Analysis was performed using a Jeol 120 EX transmission electron microscope and images were acquired with a digital camera (AMT) at 75,000–120,000 \times magnification.

Mycobacterial Protein Extracts. *M. smegmatis* and *M. tuberculosis* cells were harvested at day 7 or at 3 wk, respectively, and the culture mediums were filtrated twice on 0.22- μ m membrane and

concentrated at the desired protein concentration for proteomics analysis using Vivaspin 5-k ultrafiltration devices (SartoriusStedim Biotech) of appropriate volume.

rFasC Purification and Carbohydrate Analysis. rFasC was purified from 100 mL of 0.22 μ M filtrated culture medium (Middlebrook 7H9 broth supplemented with 0.2% glycerol) of 10 d old *M. smegmatis* mc²155:Msme_g 5996 (His-Tagged) culture using an AKTA FPLC purification system equipped with a 1 mL HisTrapHP column (GE Healthcare). rFasC jack bean α -exomannosidase (Sigma) treatment and CZE/LIF carbohydrate analysis were performed as previously described (11).

Gel Electrophoresis. Analytical and preparative (for proteomics analysis) gel electrophoresis were performed using 10 μ g and 100 μ g equivalent protein/lane on 12% (vol/vol) hand-cast analytical SDS-polyacrylamide gels (0.75 or 1.5 mm thick, 8.3 \times 7.3 cm) with a Mini-PROTEAN 3 electrophoresis system (Bio-Rad). Samples were run at a constant 20 mA for 1 h in 25 mM Tris, 192 mM glycine, and 0.1% SDS and the gels were developed with InstantBlue (Expedeon) according to the manufacturer's recommendation. The masses of the molecular standards are those reported by the manufacturer.

Mass Spectrometry Analysis. Gel lane was cut into 12 homogenous slices and treated as previously described (12). The peptides mixtures were analyzed by nanoLC-MS/MS using an Ultimate3000 system (Dionex) coupled to an LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific) for collision-induced dissociation (CID) experiments or an electron transfer dissociation (ETD)-enabled LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) for ETD experiments. The peptides were loaded on a 300- μ m inner diameter \times 5-mm PepMap C18 precolumn (LC Packings, Dionex) at 20 μ L/min in 2% (vol/vol) acetonitrile, 0.05% trifluoroacetic acid. After desalting for 5 min, peptides were separated on line on a 75- μ m inner diameter \times 15-cm C18 column (packed in-house with Reprosil C18-AQ Pur 3- μ m resin; Cluzeau). Peptides were eluted using a 5–50% (vol/vol) gradient of solvent B during 80 min at 300 nL/min flow rate. The LTQ-Orbitrap was operated in data dependent acquisition mode with the XCalibur software. Survey scan MS were acquired in the Orbitrap on the 300–2000 m/z range with the resolution set to a value of 60,000. The 20 most intense ions per survey scan were selected for both CID and ETD fragmentations and the resulting fragments were analyzed in the linear ion trap (LTQ). The normalized collision energy was set to 35% for both CID and ETD and activation times were 10 ms and 150 ms for CID and ETD fragmentation, respectively. Supplemental activation of the charge-reduced species was used in the ETD analysis to improve fragmentations. Dynamic exclusion was used within 60 s to prevent repetitive selection of the same peptide.

Data Search and Data Analysis. The Mascot Daemon software (v2.2.0, Matrix Science) was used to perform database searches in batch mode with all of the raw files acquired on each gel lane. A peak-list was created for each fraction analyzed (i.e., gel slice) and individual Mascot searches were performed for each fraction. Data were searched against actinobacteria entries in Sprot-Trembl_20090707 database (469,247 sequences). Carbamidomethylation was set as fixed modification of cysteines; oxidation of methionines and hexoses on serine and threonine were set as variable modifications for all Mascot searches. Specificity of trypsin digestion was set for cleavage after K or R, and two missed trypsin cleavages site was allowed. The mass tolerances in MS and MS/MS were set to 5 ppm and 0.8 Da, respectively, and the instrument setting was specified as "ESI-Trap." Mascot results were parsed with the in-house developed software MFPaQ v4.0.0 (Mascot File Parsing and Quantification) (13) and protein hits

were automatically validated if they satisfied one of the following criteria: identification with at least one top ranking peptide with a Mascot score of more than 47 (p -value < 0.01), or at least two top-ranking peptides each with a Mascot score of more than 33 (P < 0.05). MFPaQ was used to generate a unique nonredundant list of proteins from all of the validated result files corresponding to each fraction of the 1D gel lane. Data search and analysis was done following the guidelines edited in proteomics journals (14).

Neutral Loss Pattern Search. The detection of peak signatures corresponding to hexose loss in CID MS/MS spectra has been performed for the entire dataset using an in-house developed Perl script. The automated procedure is based on the following knowledge: the number of observed neutral losses from the fragmented parent ion is directly correlated to the number of hexoses, and such observed ions are usually of the most intense ones.

Cell Culture. Human blood samples, obtained from the Etablissement Français du Sang de Toulouse (France), were collected from nontuberculous control donors with written informed consent according to the French Public Health Code laws L1211-1, L1211-2 and L1122. Peripheral blood monocytes were centrifuged through Ficoll-Hypaque and purified by MACS separation kit and cultured to differentiation for 7 d on sterile glass coverslips in 24-well tissue culture plates (3×10^5 cells/well) containing RPMI 1640 (Gibco) supplemented with 2 mM glutamine (Gibco), 10% (vol/vol) heat inactivated FBS, penicillin-streptomycin (100 U/mL), 1% sodium pyruvate, 0.1% β -mercaptoethanol, MCSF (20 ng/mL). The culture medium was renewed on the fourth day. Macrophages derived from monocytes (MDMs) were washed twice with fresh RPMI medium before use.

The murine alveolar macrophage cell line, MH-S (ATCC; CRL-2019), was maintained in RPMI 1640 (Gibco) supplemented with 2 mM glutamine (Gibco), 10% (vol/vol) heat inactivated FBS, penicillin-streptomycin (100 U/mL), 1% sodium pyruvate, 0.1% β -mercaptoethanol, and 1% Hepes at 37 °C under an atmosphere containing 5% CO₂. Cells were plated into 24 culture plates (2.5×10^5 per well) containing a sterile glass coverslip. Before use, cells were washed twice with fresh medium.

Infections of Macrophages. Macrophage infections were performed as previously described (15). Briefly, macrophages were infected with GFP-expressing mycobacteria (16) at the multiplicity of infection of 1 in RPMI 1640. Infection was allowed to proceed at 37 °C for 1 h, and the extracellular bacteria were removed by three successive washes with fresh medium. At the end of the infection period, the cells were collected for 0, 24, 96 h, and 7 d and fixed for 1 h at room temperature with 3.7% (vol/vol) paraformaldehyde in PBS supplemented with 15 mM sucrose, and aldehyde groups were neutralized with 50 mM NH₄Cl. The cells were then rinsed in PBS, and extracellular mycobacteria were stained with rabbit anti-*Mycobacterium* Ab (1:100), which was detected by a Rhodamine red-conjugated goat anti-rabbit secondary Ab. As the cells were not permeabilized, immunostained extracellular mycobacteria displayed red and green fluorescence and intracellular bacteria appeared as green fluorescent particles. Preparations were visualized under a Leica DMLA fluorescence microscope to count and calculate (i) the number of bacteria per macrophages and (ii) the number of macrophages ingesting at least one particle (percent internalization or phagocytosis). Data are expressed as relative values, calculated as the percentage of internalization or phagocytosis divided by control values. For each set of conditions, duplicate experiments were performed and at least 100 cells per slide were counted (Table S5). The data available for the two experiments have been analyzed independently by one-way ANOVA test and t test.

Mouse Infection. Animal studies were conducted following the Centre National de la Recherche Scientifique guidelines for housing and care of laboratory animals. All protocols were reviewed and approved by the Comité d'Éthique Midi-Pyrénées (reference MP/10/12/02/11). Female 7–8 wk old SCID mice were purchased from Charles River Laboratory and housed in the Institut de Pharmacologie et de Biologie Structurale ASB3 animal facilities.

Mice, nine mice per group, were infected intranasally with 40 μ L PBS containing 10^3 cfu of either WT *M. tuberculosis* H37Rv strain or the complemented strain. For the H37Rv Δ Rv1002c mutant, three different inoculum size (10^3 , 10^4 , or 10^5 cfu) were used. Infected mice were monitored daily and animals showing signs of illness (ruffled fur, poor mobility, hunched back, or labored breathing) were killed.

1. Pelicic V, et al. (1997) Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 94(20):10955–10960.
2. Malaga W, Perez E, Guilhot C (2003) Production of unmarked mutations in mycobacteria using site-specific recombination. *FEMS Microbiol Lett* 219(2):261–268.
3. Snapper SB, Melton RE, Mustafa S, Kieser T, Jacobs WR, Jr. (1990) Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis*. *Mol Microbiol* 4(11):1911–1919.
4. Cole ST, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393(6685):537–544.
5. Bardarov S, et al. (2002) Specialized transduction: An efficient method for generating marked and unmarked targeted gene disruptions in *Mycobacterium tuberculosis*, *M. bovis* BCG and *M. smegmatis*. *Microbiology* 148(Pt 10):3007–3017.
6. Stover CK, et al. (1991) New use of BCG for recombinant vaccines. *Nature* 351(6326):456–460.
7. Simeone R, Constant P, Guilhot C, Daffé M, Chalut C (2007) Identification of the missing trans-acting enoyl reductase required for phthiocerol dimycoserate and phenolglycolipid biosynthesis in *Mycobacterium tuberculosis*. *J Bacteriol* 189(13):4597–4602.
8. Korduláková J, et al. (2002) Definition of the first mannosylation step in phosphatidylinositol mannoside synthesis. PimA is essential for growth of mycobacteria. *J Biol Chem* 277(35):31335–31344.
9. Le Dantec C, Winter N, Gicquel B, Vincent V, Picardeau M (2001) Genomic sequence and transcriptional analysis of a 23-kilobase mycobacterial linear plasmid: Evidence for horizontal transfer and identification of plasmid maintenance systems. *J Bacteriol* 183(7):2157–2164.
10. Hansen MB, Nielsen SE, Berg K (1989) Re-examination and further development of a precise and rapid dye method for measuring cell growth/cell kill. *J Immunol Methods* 119(2):203–210.
11. Ragas A, Roussel L, Puzo G, Rivière M (2007) The *Mycobacterium tuberculosis* cell-surface glycoprotein apa as a potential adhesin to colonize target cells via the innate immune system pulmonary C-type lectin surfactant protein A. *J Biol Chem* 282(8):5133–5142.
12. Mouton-Barbosa E, et al. (2010) In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. *Mol Cell Proteomics* 9(5):1006–1021.
13. Bouyssie D, et al. (2007) Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: Application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol Cell Proteomics* 6(9):1621–1637.
14. Bradshaw RA, Burlingame AL, Carr S, Aebersold R (2006) Reporting protein identification data: The next generation of guidelines. *Mol Cell Proteomics* 5(5):787–788.
15. Astarie-Dequeker C, et al. (1999) The mannose receptor mediates uptake of pathogenic and nonpathogenic mycobacteria and bypasses bactericidal responses in human macrophages. *Infect Immun* 67(2):469–477.
16. Astarie-Dequeker C, et al. (2009) Phthiocerol dimycoserates of *M. tuberculosis* participate in macrophage invasion by inducing changes in the organization of plasma membrane lipids. *PLoS Pathog* 5(2):e1000289.

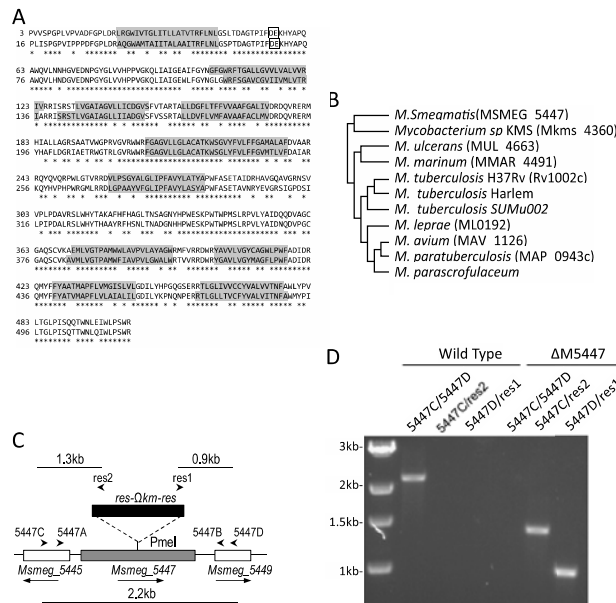


Fig. S1. (A) Sequence alignment of the *M. tuberculosis* (Mtb) Rv1002c (Top sequence) and *M. smegmatis* Msmeg_5447 putative gene products showing the catalytic (DE amino acid pair (squared) and the hydrophobic segments (shadowed) assessing the predicted transmembrane location of the protein mannitol transferases (PMT). (B) Dendrogram of the mycobacterial homologs of the Mtb PMT encoded by the Rv1002c gene. (C) Schematic diagram of the genomic organization of the Msmeg_5447 locus depicting the targeted (gray box) and its adjacent genes (white boxes) and the K_m resistance cassette (black box). Positions of primers used for the characterization of the mutant strain are indicated by arrowheads and the expected sizes for PCR products are indicated. (D) PCR analysis of the WT and JM5447 mutant using various combinations of specific primers.

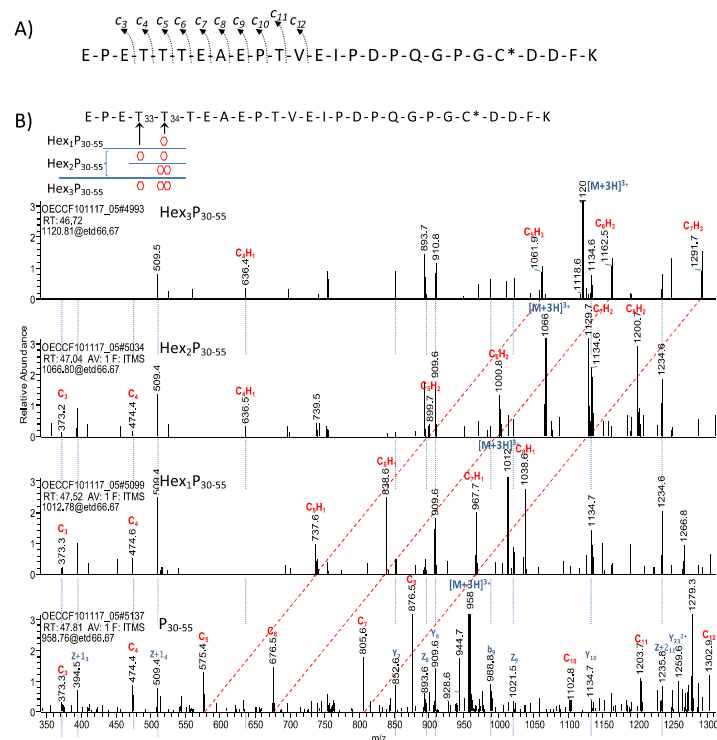


Fig. S2. MS/MS glycosylation site determination of the rFasC N-terminal glycopeptide P₃₀₋₅₅. (A) rFasC N-terminal peptide P₃₀₋₅₅ sequence reporting the major peptide bound cleavages giving rise to the informative peptide fragment ion series observed in the positive ETD MS/MS spectra. (B) ETD MS/MS analysis of the triply charged precursor ions of the different glycoforms of the rFasC P₃₀₋₅₅ peptide (at *m/z* 958.76: unglycosylated, *m/z* 1012.78: monoglycosylated, *m/z* 1066.80: diglycosylated and *m/z* 1120.81: triglycosylated peptide). Localization and occupation of the glycosylation sites can be deduced from the mass shifts of the C-fragment ion series observed in the ETD MS/MS spectra of the different P₃₀₋₅₅ glycoforms reported in Table S3 and are reported in the *Inset*. (Asterisk indicates a propionamide modified cysteine).

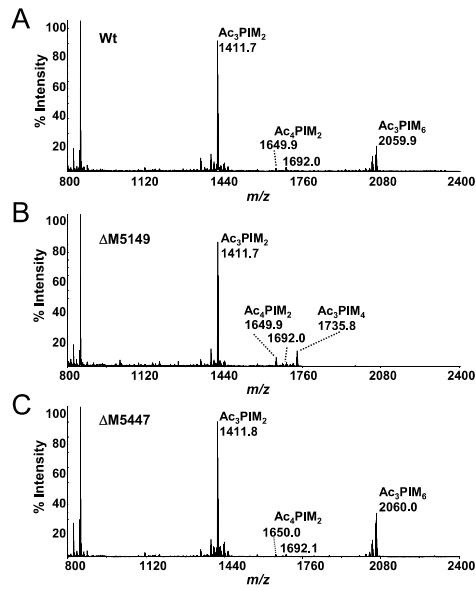


Fig. S3. Mass spectrometry analysis of the impact of the inactivation of the PimE mannosyl transferase (*Msmeg_5149*) and the PMT (*Msmeg_5447*) on the *M. smegmatis* phosphatidyl inositol mannoside (PIM) glycolipids biosynthesis evidencing the extinction of the polar PIM lineage (PIM6) in the PimE invalidated mutant Δ M5149 (**B**) and the absence of effect of the PMT gene interruption on the PIM glycolipids synthesis (**C**). PIM mixtures, prepared as previously described from cultures of *M. smegmatis* WT (**A**), PimE (**B**), and PMT mutant (**C**), were analyzed by MALDI-TOF MS in negative-ion detection mode (1).

1. Cala-de Paepe D, et al. (2012) Deciphering the role of CD1e protein in mycobacterial phosphatidyl-myo-inositol mannosides (PIM) processing for presentation by CD1b to T lymphocytes. *J Biol Chem* 287(37):31494–31502.

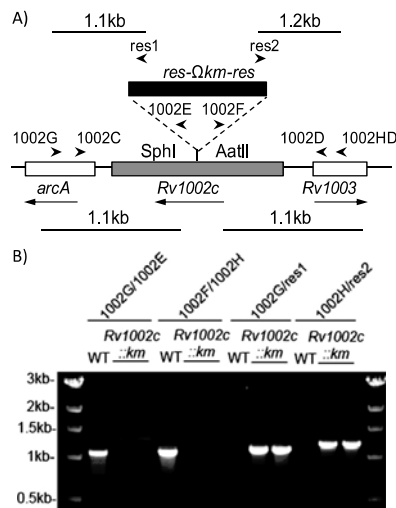


Fig. 54. (A) Schematic diagram of the genomic organization of the *Rv1002c* locus (see legend to Fig. S1C for symbol descriptions). (B) PCR analysis of the WT and the Mtb $\Delta Rv1002c$ mutant (*Rv1002c::km*) using various combinations of specific primers.

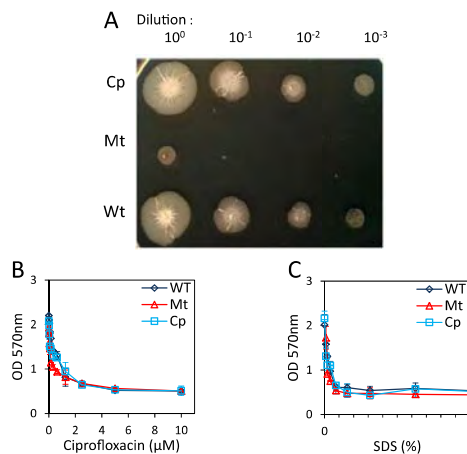


Fig. 55. (A) Growth of the Mtb strains on Malachite green free 7H9 agar solid culture medium evidencing that the mutant growth defect associated to the *Rv1002c* protein-*O*-mannosyl transferase gene disruption is not because of an increased sensitivity of the mutant to the bacteriostatic property of the Malachite green. (B and C) Mtb strains sensitivity to the antituberculosis drug Ciprofloxacin (B), or to cell wall stress induced by detergent (SDS) (C). One-hundred microliters of bacterial suspension (OD ~0.04) of Mtb WT, $\Delta Rv1002c$ (Mt), or complemented mutant (Cp) strains were incubated on 7H9 + 0.05% Tween80 and 2.5% (vol/vol) ADC for 6 d in the absence or presence of increasing concentration of Ciprofloxacin (0–10 μM) or SDS (0–1%) and survival was followed by monitoring the level of MTT reduction by the metabolically active cells at 570 nm.

Table S4. Name and main features of strain, plasmids, and primers used in this study

Name	Relevant characteristics	Trivial name	Source
Strains			
H37Rv	<i>M. tuberculosis</i>	Wild-type	(4)
mc ² 155	<i>M. smegmatis</i>	Wild-type	(3)
PMM143	<i>M. smegmatis</i> mc ² 155 <i>Msmeg_5447::res-Ωkm-res</i>	ΔM5447	Present study
PMM152	<i>M. tuberculosis</i> H37Rv <i>Rv1002c::res-Ωkm-res</i>	ΔRv1002c	Present study
PMM176	<i>M. tuberculosis</i> H37Rv <i>Rv1002c::res</i>	ΔRv1002c	Present study
	<i>M. smegmatis</i> mc ² 155 <i>Msmeg_4247::km</i>	ΔM4247	Present study
	<i>M. smegmatis</i> mc ² 155 <i>Msmeg_5149::km</i>	ΔM5149	Present study
Plasmids and phages			
pYUB854	Cosmid vector containing λ <i>cos</i> site		(5)
phAE87	Shuttle phasmid derived from mycobacteriophage TM4		(5)
phWM11	Shuttle phasmid derived from phAE87 and containing <i>Rv1002c::res-Ωkm-res</i>		Present study
pWM19	Thermosensitive <i>E. coli</i> /mycobacteria shuttle plasmid containing the resolvase gene from transposon γδ		(2)
pMV361	Integrative <i>Escherichia coli</i> /mycobacteria shuttle vector		(6)
pMIP12	Replicative <i>E. coli</i> /mycobacteria shuttle plasmid		(9)
pVM158H	Mycobacterial plasmid derived from pMV361 containing the <i>Msmeg_5447</i> gene under the control of the <i>pBlaF*</i> promoter		Present study
pWM159H	Mycobacterial plasmid derived from pMV361 containing the <i>Rv1002c</i> gene under the control of the <i>pBlaF*</i> promoter		Present study
pWM218	Mycobacterial plasmid derived from pMIP12 containing the <i>FasC</i> gene under the control of the <i>pBlaF*</i> promoter		Present study
pWM292	Mycobacterial plasmid derived from pMV361 containing the <i>Msmeg_5149</i> gene under the control of the <i>pBlaF*</i> promoter		Present study
pVVRv1159	Mycobacterial plasmid derived from pVV16 (8) containing the <i>Rv1002c</i> gene under the control of the <i>phsp60</i> promoter.		Present study
pCG211	Mycobacterial plasmid derived from pMIP12 containing the <i>gfp</i> gene		(16)
Primers			
res1	5'- TAACAGACACTGCTGTCCGATA-3'		
res2	5'- TATCGGACAAGCAGTGTCTGTTA-3'		
5447A	5'- ATAGATATCGCAGAAAGCGATCCGACGC-3'		
5447B	5'- ATAGATATCCTGTGGTGTACGGTTTCG-3'		
5447C	5'- GCGTCTGAAGCCATGTCATG-3'		
5447D	5'- GGGTCAGTACCACCTTGCAG3-3'		
5447E	5'- ATCATATGACCGCCCTCGACAC-3'		
5447F	5'- ATAAGCTTCTAGCGCCAGCTCGGCA-3'		
5196A	5'- ATCATATGCACACGCGTACCAAGAC-3'		
5196B	5'- ATACTAGTGCCGATCGGGGCGTCGG-3'		
1002A	5'- ATCATATGGTACCCGTCGTCAGCC-3'		
1002B	5'- ATAAGCTTACCCGCGAGTGGGCA-3'		
1002C	5'- ATGCTAGCCACGTTGTCGACACGC-3'		
1002D	5'- ATGCTAGCACCGTCGACCGTTATT-3'		
1002E	5'- GTCCGAGACCGCATCG-3'		
1002F	5'- AGCCACGGTCTTGCACCTTG-3'		
1002G	5'- GCATCTGGACACGGTGTGC-3'		
1002H	5'- CGGGATCGGTGATACCG-3'		

Liu et al. www.pnas.org/cgi/content/short/1219704110

Because of incomplete dataset for exp. 2, the graphs of Fig. 4 correspond to the exp. 1 dataset. Percent of infection = no. of infected cells \times 100/no. of counted cells. Bacterial load = no. of total bacterial/no. of infected cells.

Because of incomplete dataset for exp. 2, the graphs of Fig. 4 correspond to the exp. 1 dataset. Percent of infection = no. of infected cells \times 100/no. of counted cells. Bacterial load = no. of total bacterial/no. of infected cells.

Bibliographie

- Abu-Qarn, M., J. Eichler and N. Sharon (2008). "Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea." Curr Opin Struct Biol **18**(5): 544-550.
- Ahlf, D. R., P. M. Thomas and N. L. Kelleher (2013). "Developing top down proteomics to maximize proteome and sequence coverage from cells and tissues." Curr Opin Chem Biol **17**(5): 787-794.
- Almeida Da Silva, P. E. and J. C. Palomino (2011). "Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs." J Antimicrob Chemother **66**(7): 1417-1430.
- Alpert, A. J. (1990). "Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds." Journal of Chromatography A **499**(0): 177-196.
- Alvarez-Manilla, G., J. Atwood, 3rd, Y. Guo, N. L. Warren, R. Orlando and M. Pierce (2006). "Tools for glycoproteomic analysis: size exclusion chromatography facilitates identification of tryptic glycopeptides with N-linked glycosylation sites." J Proteome Res **5**(3): 701-708.
- Andersen, P. and T. M. Doherty (2005). "The success and failure of BCG [mdash] implications for a novel tuberculosis vaccine." Nat Rev Micro **3**(8): 656-662.
- Andries, K., P. Verhasselt, J. Guillemont, H. W. H. Göhlmann, J.-M. Neefs, H. Winkler, J. Van Gestel, P. Timmerman, M. Zhu, E. Lee, P. Williams, D. de Chaffoy, E. Huitric, S. Hoffner, E. Cambau, C. Truffot-Pernot, N. Lounis and V. Jarlier (2005). "A Diarylquinoline Drug Active on the ATP Synthase of Mycobacterium tuberculosis." Science **307**(5707): 223-227.
- Antberg, L., P. Cifani, M. Sandin, F. Levander and P. James (2012). "Critical comparison of multidimensional separation methods for increasing protein expression coverage." J Proteome Res **11**(5): 2644-2652.
- Bandhakavi, S., S. K. Van Riper, P. N. Tawfik, M. D. Stone, T. Haddad, N. L. Rhodus, J. V. Carlis and T. J. Griffin (2011). "Hexapeptide libraries for enhanced protein PTM identification and relative abundance profiling in whole human saliva." J Proteome Res **10**(3): 1052-1061.
- Bantscheff, M., S. Lemeer, M. M. Savitski and B. Kuster (2012). "Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present." Anal Bioanal Chem **404**(4): 939-965.
- Becker, J. W., G. N. Reeke, Jr., J. L. Wang, B. A. Cunningham and G. M. Edelman (1975). "The covalent and three-dimensional structure of concanavalin A. III. Structure of the monomer and its interactions with metals and saccharides." J Biol Chem **250**(4): 1513-1524.
- Bell, C., G. T. Smith, M. J. Sweredoski and S. Hess (2012). "Characterization of the Mycobacterium tuberculosis proteome by liquid chromatography mass spectrometry-based proteomics techniques: a comprehensive resource for tuberculosis research." J Proteome Res **11**(1): 119-130.

Biemann, K. (1990). "Appendix 5. Nomenclature for peptide fragment ions (positive ions)." Methods Enzymol **193**: 886-887.

Bildl, W., A. Haupt, C. S. Muller, M. L. Biniossek, J. O. Thumfart, B. Huber, B. Fakler and U. Schulte (2012). "Extending the dynamic range of label-free mass spectrometric quantification of affinity purifications." Mol Cell Proteomics **11**(2): M111 007955.

Bondarenko, P. V., T. P. Second, V. Zabrouskov, A. A. Makarov and Z. Zhang (2009). "Mass measurement and top-down HPLC/MS analysis of intact monoclonal antibodies on a hybrid linear quadrupole ion trap-Orbitrap mass spectrometer." J Am Soc Mass Spectrom **20**(8): 1415-1424.

Bourzac, K. (2014). "Infectious disease: Beating the big three." Nature **507**(7490): S4-S7.

Bouyssie, D., A. Gonzalez de Peredo, E. Mouton, R. Albigot, L. Roussel, N. Ortega, C. Cayrol, O. Burlet-Schiltz, J. P. Girard and B. Monsarrat (2007). "Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells." Mol Cell Proteomics **6**(9): 1621-1637.

Brosch, R., S. V. Gordon, T. Garnier, K. Eiglmeier, W. Frigui, P. Valenti, S. Dos Santos, S. Duthoy, C. Lacroix, C. Garcia-Pelayo, J. K. Inwald, P. Golby, J. N. Garcia, R. G. Hewinson, M. A. Behr, M. A. Quail, C. Churcher, B. G. Barrell, J. Parkhill and S. T. Cole (2007). "Genome plasticity of BCG and impact on vaccine efficacy." Proc Natl Acad Sci U S A **104**(13): 5596-5601.

Brun, V., A. Dupuis, A. Adrait, M. Marcellin, D. Thomas, M. Court, F. Vandenesch and J. Garin (2007). "Isotope-labeled protein standards: toward absolute quantitative proteomics." Mol Cell Proteomics **6**(12): 2139-2149.

Carapito, C., C. Klemm, R. Aebersold and B. Domon (2009). "Systematic LC-MS analysis of labile post-translational modifications in complex mixtures." J Proteome Res **8**(5): 2608-2614.

Carmona, J., A. Cruz, L. Moreira-Teixeira, C. Sousa, J. Sousa, N. S. Saraiva, A. L. Svenson, G. Kallenius, J. Pedrosa, F. Rodrigues, A. G. Castro and M. Saraiva (2013). "Strains are differentially recognized by TLRs with an impact on the immune response." PLoS One, **8**.

Chalkley, R. J., P. R. Baker, K. F. Medzihradszky, A. J. Lynn and A. L. Burlingame (2008). "In-depth analysis of tandem mass spectrometry data from disparate instrument types." Mol Cell Proteomics **7**(12): 2386-2398.

Chatterjee, D., K. Lowell, B. Rivoire, M. R. McNeil and P. J. Brennan (1992). "Lipoarabinomannan of Mycobacterium tuberculosis. Capping with mannosyl residues in some strains." J Biol Chem **267**(9): 6234-6239.

Chauhan, S., D. Sharma, A. Singh, A. Surolia, J.S. Tyagi (2011). "Comprehensive insights into Mycobacterium tuberculosis DevR (DosR) regulon activation switch." Nucleic Acids Res **39**(17): 7400-7414.

Clemmer, D. E. and M. F. Jarrold (1997). "Ion Mobility Measurements and their Applications to Clusters and Biomolecules." Journal of Mass Spectrometry **32**(6): 577-592.

Collier, T. S., S. M. Randall, P. Sarkar, B. M. Rao, R. A. Dean and D. C. Muddiman (2011). "Comparison of stable-isotope labeling with amino acids in cell culture and spectral counting for relative quantification of protein expression." Rapid Commun Mass Spectrom **25**(17): 2524-2532.

Coon, J. J., B. Ueberheide, J. E. Syka, D. D. Dryhurst, J. Ausio, J. Shabanowitz and D. F. Hunt (2005). "Protein identification using sequential ion/ion reactions and tandem mass spectrometry." Proc Natl Acad Sci U S A **102**(27): 9463-9468.

Cox, J. and M. Mann (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." Nat Biotechnol **26**(12): 1367-1372.

Cox, K. A., S. J. Gaskell, M. Morris and A. Whiting (1996). "Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions." J Am Soc Mass Spectrom **7**(6): 522-531.

Daniel, T. M. (2005). "Robert Koch and the pathogenesis of tuberculosis." Int J Tuberc Lung Dis **9**(11): 1181-1182.

Daniel, T. M. (2006). "The history of tuberculosis." Respir Med **100**(11): 1862-1870.

Davids, V., W. A. Hanekom, N. Mansoor, H. Gamielien, S. J. Gelderbloem, A. Hawkrige, G. D. Hussey, E. J. Hughes, J. Soler, R. A. Murray, S. R. Ress and G. Kaplan (2006). "The effect of bacille Calmette-Guerin vaccine strain and route of administration on induced immune responses in vaccinated infants." J Infect Dis **193**(4): 531-536.

de Souza, G. A. and H. G. Wiker (2011). "A proteomic view of mycobacteria." Proteomics **11**(15): 3118-3127.

Debaene, F., E. Wagner-Rousset, O. Colas, D. Ayoub, N. Corvaia, A. Van Dorsselaer, A. Beck and S. Cianfèrani (2013). "Time Resolved Native Ion-Mobility Mass Spectrometry to Monitor Dynamics of IgG4 Fab Arm Exchange and "Bispecific" Monoclonal Antibody Formation." Analytical Chemistry **85**(20): 9785-9792.

Dobos, K. M., K. H. Khoo, K. M. Swiderek, P. J. Brennan and J. T. Belisle (1996). "Definition of the full extent of glycosylation of the 45-kilodalton glycoprotein of Mycobacterium tuberculosis." J Bacteriol **178**(9): 2498-2506.

Dole, M. W., L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson and M. B. Alice (1968). "Molecular beams of macroions." J. Chem. Phys **49**(2240).

Domon, B. and R. Aebersold (2006). "Mass spectrometry and protein analysis." Science **312**(5771): 212-217.

Domon, B. and C. Costello (1988). "A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates." Glycoconjugate Journal **5**(4): 397-409.

Elias, J. E. and S. P. Gygi (2010). "Target-decoy search strategy for mass spectrometry-based proteomics." Methods Mol Biol **604**: 55-71.

Espitia, C. and R. Mancilla (1989). "Identification, isolation and partial characterization of Mycobacterium tuberculosis glycoprotein antigens." Clin Exp Immunol **77**(3): 378-383.

Eyers, C. E., C. Lawless, D. C. Wedge, K. W. Lau, S. J. Gaskell and S. J. Hubbard (2011). "CONSequence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches." Mol Cell Proteomics **10**(11): M110.003384.

Gonzalez-Zamorano M., G. Mendoza-Hernandez, W. Xolalpa, C. Parada, A.J. Vallecillo, F. Bigi and C. Espitia (2009). "Mycobacterium tuberculosis glycoproteomics based on ConA-lectin affinity capture of mannosylated proteins." Journal of Proteome Research **8**(2): 721-33.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse (1989). "Electrospray ionization for mass spectrometry of large biomolecules." Science **246**(4926): 64-71.

Fine, P. E. (1995). "Variation in protection by BCG: implications of and for heterologous immunity." Lancet **346**(8986): 1339-1345.

Fornelli, L., E. Damoc, P. M. Thomas, N. L. Kelleher, K. Aizikov, E. Denisov, A. Makarov and Y. O. Tsybin (2012). "Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS." Mol Cell Proteomics **11**(12): 1758-1767.

Frank, A. M., J. J. Pesavento, C. A. Mizzen, N. L. Kelleher and P. A. Pevzner (2008). "Interpreting top-down mass spectra using spectral alignment." Anal Chem **80**(7): 2499-2505.

Frese, C. K., H. Zhou, T. Taus, A. F. M. Altelaar, K. Mechtler, A. J. R. Heck and S. Mohammed (2013). "Unambiguous Phosphosite Localization using Electron-Transfer/Higher-Energy Collision Dissociation (ETHcD)." Journal of Proteome Research **12**(3): 1520-1525.

Gabelica, V. (2006). "Intérêt de la spectrométrie de mobilité ionique pour l'étude de la conformation et des assemblages non-covalents de biomolécules." Spectra analyse **251**: 21-27.

Garcia, B. A., J. J. Pesavento, C. A. Mizzen and N. L. Kelleher (2007). "Pervasive combinatorial modification of histone H3 in human cells." Nat Meth **4**(6): 487-489.

Gault, J., C. Malosse, G. Dumenil and J. Chamot-Rooke (2013). "A combined mass spectrometry strategy for complete posttranslational modification mapping of Neisseria meningitidis major pilin." J Mass Spectrom **48**(11): 1199-1206.

Gautier, V., E. Mouton-Barbosa, D. Bouyssie, N. Delcourt, M. Beau, J.P. Girard, C. Cayrol, O. Burlet-Schiltz, B. Monsarrat, A. Gonzalez de Peredo (2012). "Label-free quantification and shotgun analysis of complex proteomes by one-dimensional SDS-PAGE/NanoLC-MS: evaluation for the large scale analysis of inflammatory human endothelial cells." Mol Cell Proteomics **11**(8): 527-539.

Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant (2004). "Open mass spectrometry search algorithm." J Proteome Res **3**(5): 958-964.

Geijtenbeek, T. B., S. J. Van Vliet, E. A. Koppel, M. Sanchez-Hernandez, C. M. Vanderbroucke-Grauls, B. Appelmek and Y. Van Kooyk (2003). "Mycobacteria target DC-SIGN to suppress dendritic cell function." J Exp Med **197**: 7-17.

Gerber, S. A., J. Rush, O. Stemman, M. W. Kirschner and S. P. Gygi (2003). "Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS." Proc Natl Acad Sci U S A **100**(12): 6940-6945.

Gilar, M., A. Jaworski, P. Olivova and J. C. Gebler (2007). "Peptide retention prediction applied to proteomic data analysis." Rapid Commun Mass Spectrom **21**(17): 2813-2821.

Gooley, A. A., B. J. Classon, R. Marschalek and K. L. Williams (1991). "Glycosylation sites identified by detection of glycosylated amino acids released from Edman degradation: the identification of Xaa-Pro-Xaa-Xaa as a motif for Thr-O-glycosylation." Biochem Biophys Res Commun **178**(3): 1194-1201.

Guerrier, L., V. Thulasiraman, A. Castagna, F. Fortis, S. Lin, L. Lomas, P. G. Righetti and E. Boschetti (2006). "Reducing protein concentration range of biological samples using solid-phase ligand libraries." J Chromatogr B Analyt Technol Biomed Life Sci **833**(1): 33-40.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**(10): 994-999.

Haginaka, J., H. Tabo and H. Matsunaga (2012). "Preparation of molecularly imprinted polymers for organophosphates and their application to the recognition of organophosphorus compounds and phosphopeptides." Anal Chim Acta **748**: 1-8.

Hao, Z., L. Jiang, I. Mohtashemi, A. Huhmer (2010). "N- and C-terminal sequencing of proteins using top-down electron transfer dissociation mass spectrometry". Thermo application note **484**.

Herrmann, J. L., R. Delahay, A. Gallagher, B. Robertson and D. Young (2000). "Analysis of post-translational modification of mycobacterial proteins using a cassette expression system." FEBS Lett **473**(3): 358-362.

Herrmann, J. L., P. O'Gaora, A. Gallagher, J. E. Thole and D. B. Young (1996). "Bacterial glycoproteins: a link between glycosylation and proteolytic cleavage of a 19 kDa antigen from Mycobacterium tuberculosis." EMBO J **15**(14): 3547-3554.

Hoffmann, C., A. Leis, M. Niederweis, J. M. Plitzko and H. Engelhardt (2008). "Disclosure of the mycobacterial outer membrane: Cryo-electron tomography and vitreous sections reveal the lipid bilayer structure." Proceedings of the National Academy of Sciences **105**(10): 3963-3967.

Horn, C., A. Namane, P. Pescher, M. Riviere, F. Romain, G. Puzo, O. Barzu and G. Marchal (1999). "Decreased capacity of recombinant 45/47-kDa molecules (Apa) of Mycobacterium tuberculosis to stimulate T lymphocyte responses related to changes in their mannosylation pattern." J Biol Chem **274**(45): 32023-32030.

Hunter, S. W. and P. J. Brennan (1990). "Evidence for the presence of a phosphatidylinositol anchor on the lipoarabinomannan and lipomannan of *Mycobacterium tuberculosis*." J Biol Chem **265**(16): 9272-9279.

Janeway, C. A., Jr. and R. Medzhitov (2002). "Innate immune recognition." Annu Rev Immunol **20**: 197-216.

Jayakumar, D., W. R. Jacobs, Jr. and S. Narayanan (2008). "Protein kinase E of *Mycobacterium tuberculosis* has a role in the nitric oxide stress response and apoptosis in a human macrophage model of infection." Cell Microbiol **10**(2): 365-374.

Julenius, K., A. Molgaard, R. Gupta and S. Brunak (2005). "Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites." Glycobiology **15**(2): 153-164.

Kahn, P. (1995). "From genome to proteome: looking at a cell's proteins." Science **270**(5235): 369-370.

Kang, P. B., A. K. Azad, J. B. Torrelles, T. M. Kaufman, A. Beharka, E. Tibesar, L. E. Desjardin and L. S. Schlesinger (2005). "The human macrophage mannose receptor directs *Mycobacterium tuberculosis* lipoarabinomannan-mediated phagosome biogenesis." J Exp Med **202**: 987-999.

Kebarle, P. (2000). "A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry." J Mass Spectrom **35**(7): 804-817.

Kleinnijenhuis, J., M. Oosting, L. A. Joosten, M. G. Netea and R. Van Crevel (2011). "Innate immune recognition of *Mycobacterium tuberculosis*." Clinical and development immunology.

Kordulakova, J., M. Gilleron, K. Mikusova, G. Puzo, P. J. Brennan, B. Gicquel and M. Jackson (2002). "Definition of the first mannosylation step in phosphatidylinositol mannoside synthesis. PimA is essential for growth of mycobacteria." J Biol Chem **277**(35): 31335-31344.

Kuster, B., M. Schirle, P. Mallick and R. Aebersold (2005). "Scoring proteomes with proteotypic peptide probes." Nat Rev Mol Cell Biol **6**(7): 577-583.

Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature **227**(5259): 680-685.

Lanucara, F. and C. E. Eyers (2013). "Top-down mass spectrometry for the analysis of combinatorial post-translational modifications." Mass Spectrom Rev **32**(1): 27-42.

Larsen, M. R., S. J. Cordwell and P. Roepstorff (2002). "Graphite powder as an alternative or supplement to reversed-phase material for desalting and concentration of peptide mixtures prior to matrix-assisted laser desorption/ionization-mass spectrometry." Proteomics **2**(9): 1277-1287.

LeDuc, R. D., G. K. Taylor, Y. B. Kim, T. E. Januszyk, L. H. Bynum, J. V. Sola, J. S. Garavelli and N. L. Kelleher (2004). "ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry." Nucleic Acids Res **32**(Web Server issue): W340-345.

Lengeler, KB., D. Tielker and J.F. Ernst (2008). "Protein-O-mannosyltransferase in virulence and development." Cell Mol Life Sci **65**(4): 528-44.

Li, F., O. V. Glinskii and V. V. Glinsky (2013). "Glycobioinformatics: current strategies and tools for data mining in MS-based glycoproteomics." Proteomics **13**(2): 341-354.

Liu, H., R. G. Sadygov and J. R. Yates, 3rd (2004). "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." Anal Chem **76**(14): 4193-4201.

Liu, X., Y. Sirotkin, Y. Shen, G. Anderson, Y. S. Tsai, Y. S. Ting, D. R. Goodlett, R. D. Smith, V. Bafna and P. A. Pevzner (2012). "Protein identification using top-down." Mol Cell Proteomics **11**(6): M111.008524.

Liu, N. Q., L. J. Dekker, C. Stingl, C. Güzel, T. De Marchi, J. W. M. Martens, J. A. Foekens, T. M. Luider and A. Umar (2013). "Quantitative proteomic analysis of microdissected breast cancer tissues: comparison of label-free and SILAC-based quantification with shotgun, directed, and targeted MS approaches." J Proteome Res **12**(10): 4627-4641.

Mahne, M., A. Tauch, A. Puhler and J. Kalinowski (2006). "The *Corynebacterium glutamicum* gene *pmt* encoding a glycosyltransferase related to eukaryotic protein-O-mannosyltransferases is essential for glycosylation of the resuscitation promoting factor (Rpf2) and other secreted proteins." FEMS Microbiol Lett **259**(2): 226-233.

Makarov, V., G. Manina, K. Mikusova, U. Mollmann, O. Ryabova, B. Saint-Joanis, N. Dhar, M. R. Pasca, S. Buroni, A. P. Lucarelli, A. Milano, E. De Rossi, M. Belanova, A. Bobovska, P. Dianiskova, J. Kordulakova, C. Sala, E. Fullam, P. Schneider, J. D. McKinney, P. Brodin, T. Christophe, S. Waddell, P. Butcher, J. Albrethsen, I. Rosenkrands, R. Brosch, V. Nandi, S. Bharath, S. Gaonkar, R. K. Shandil, V. Balasubramanian, T. Balganes, S. Tyagi, J. Grosset, G. Riccardi and S. T. Cole (2009). "Benzothiazinones kill *Mycobacterium tuberculosis* by blocking arabinan synthesis." Science **324**(5928): 801-804.

Matsumoto, M., H. Hashizume, T. Tomishige, M. Kawasaki, H. Tsubouchi, H. Sasaki, Y. Shimokawa and M. Komatsu (2006). "OPC-67683, a nitro-dihydro-imidazo[4,5-c]pyridine derivative with promising action against tuberculosis in vitro and in mice." PLoS Med **3**(11): e466.

Matsumoto, M., T. Tanaka, T. Kaisho, H. Sanjo, N. G. Copeland, D. J. Gilbert, N. A. Jenkins and S. Akira (1999). "A novel LPS-inducible C-type lectin is a transcriptional target of NF-IL6 in macrophages." J Immunol **163**(9): 5039-5048.

McDaniel, E. W., D. W. Martin and W. S. Barnes (1962). "Drift Tube-Mass Spectrometer for Studies of Low-Energy Ion-Molecule Reactions." Review of Scientific Instruments **33**(1): 2-7.

McLafferty, F. W. (1981). "Tandem mass spectrometry." Science **214**(4518): 280-287.

Medzhitov, R. (2007). "Recognition of microorganisms and activation of the immune response." Nature **449**(7164): 819-826.

Mescher, M. F. and J. L. Strominger (1976). "Purification and characterization of a procaryotic glucoprotein from the envelope of *Halobacterium salinarum*." J Biol Chem **251**(7): 2005-2014.

Michell, S. L., A. O. Whelan, P. R. Wheeler, M. Panico, R. L. Easton, A. T. Etienne, S. M. Haslam, A. Dell, H. R. Morris, A. J. Reason, J. L. Herrmann, D. B. Young and R. G. Hewinson (2003). "The MPB83 antigen from *Mycobacterium bovis* contains O-linked mannose and (1-->3)-mannobiose moieties." J Biol Chem **278**(18): 16423-16432.

Mitchison, D. and G. Davies (2012). "The chemotherapy of tuberculosis: past, present and future." Int J Tuberc Lung Dis **16**(6): 724-732.

Moradian, A., A. Kalli, M. J. Sweredoski and S. Hess (2014). "The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications." Proteomics **14**(4-5): 489-497.

Moradian, A., A. Kalli, M. J. Sweredoski and S. Hess (2014). "The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications." PROTEOMICS **14**(4-5): 489-497.

Morita, Y. S., C. B. Sena, R. F. Waller, K. Kurokawa, M. F. Sernee, F. Nakatani, R. E. Haite, H. Billman-Jacobe, M. J. McConville, Y. Maeda and T. Kinoshita (2006). "PimE is a polyprenol-phosphate-mannose-dependent mannosyltransferase that transfers the fifth mannose of phosphatidylinositol mannoside in mycobacteria." J Biol Chem **281**(35): 25143-25155.

Mostovenko, E., C. Hassan, J. Rattke, A. M. Deelder, P. A. van Veelen and M. Palmblad (2013). "Comparison of peptide and protein fractionation methods in proteomics." EuPA Open Proteomics **1**(0): 30-37.

Mueller, L. N., M. Y. Brusniak, D. R. Mani and R. Aebersold (2008). "An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data." J Proteome Res **7**(1): 51-61.

Nesvizhskii, A. I. (2010). "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics." J Proteomics **73**(11): 2092-2123.

Nigou, J., C. Zelle-Rieser, M. Gilleron, M. Thurnher and G. Puzo (2001). "Mannosylated lipoarabinomannans inhibit IL-12 production by human dendritic cells: evidence for a negative signal delivered through the mannose receptor." J Immunol, **166**: 7477-7485.

Olsen, J. V., B. Macek, O. Lange, A. Makarov, S. Horning and M. Mann (2007). "Higher-energy C-trap dissociation for peptide modification analysis." Nat Methods **4**(9): 709-712.

Olsen, J. V., S. E. Ong and M. Mann (2004). "Trypsin cleaves exclusively C-terminal to arginine and lysine residues." Mol Cell Proteomics **3**(6): 608-614.

Olsen, J. V., J. C. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E. R. Wouters, M. Senko, A. Makarov, M. Mann and S. Horning (2009). "A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed." Mol Cell Proteomics **8**(12): 2759-2769.

Ong, S. E. and M. Mann (2005). "Mass spectrometry-based proteomics turns quantitative." Nat Chem Biol **1**(5): 252-262.

Ortalo-Magne, A., M. A. Dupont, A. Lemassu, A. B. Andersen, P. Gounon and M. Daffe (1995). "Molecular composition of the outermost capsular material of the tubercle bacillus." Microbiology **141** (Pt 7): 1609-1620.

Pan, S., R. Chen, R. Aebersold, TA. Brentnall (2011). "Mass spectrometry based glycoproteomics - from a peoteomics perspective." Mol Cell Proteomics **10**(1)

Patterson, S. D. and R. H. Aebersold (2003). "Proteomics: the first decade and beyond." Nat Genet **33 Suppl**: 311-323.

Paulson, T. (2013). "Epidemiology: A mortal foe." Nature **502**(7470): S2-S3.

Perkins, D. N., D. J. Pappin, D. M. Creasy and J. S. Cottrell (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." Electrophoresis **20**(18): 3551-3567.

Picotti, P. and R. Aebersold (2012). "Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions." Nat Methods **9**(6): 555-566.

Pitarque, S., G. Larrouy-Maumus, B. Payre, M. Jackson, G. Puzo and J. Nigou (2008). "The immunomodulatory lipoglycans, lipoarabinomannan and lipomannan, are exposed at the mycobacterial cell surface." Tuberculosis (Edinb) **88**(6): 560-565.

Ragas, A., L. Roussel, G. Puzo and M. Riviere (2007). "The Mycobacterium tuberculosis cell-surface glycoprotein apa as a potential adhesin to colonize target cells via the innate immune system pulmonary C-type lectin surfactant protein A." J Biol Chem **282**(8): 5133-5142.

Raju, R. M., M. Unnikrishnan, D. H. F. Rubin, V. Krishnamoorthy, O. Kandror, T. N. Akopian, A. L. Goldberg and E. J. Rubin (2012). "Mycobacterium tuberculosis ClpP1 and ClpP2 function together in protein degradation and are required for viability in vitro and during infection". Plos Pathogens **8**(2).

Rawn, J. D. and G. E. Lienhard (1974). "The binding of boronic acids to chymotrypsin." Biochemistry **13**(15): 3124-3130.

Ritz, N., W. A. Hanekom, R. Robins-Browne, W. J. Britton and N. Curtis (2008). "Influence of BCG vaccine strain on the immune response and protection against tuberculosis." FEMS Microbiol Rev **32**(5): 821-841.

Romain, F., C. Horn, P. Pescher, A. Namane, M. Riviere, G. Puzo, O. Barzu and G. Marchal (1999). "Deglycosylation of the 45/47-kilodalton antigen complex of Mycobacterium tuberculosis decreases its capacity to elicit in vivo or in vitro cellular immune responses." Infect Immun **67**(11): 5567-5572.

Rose, R. J., E. Damoc, E. Denisov, A. Makarov and A. J. Heck (2012). "High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies." Nat Methods **9**(11): 1084-1086.

Russell, D. G., C. E. Barry, 3rd and J. L. Flynn (2010). "Tuberculosis: what we don't know can, and does, hurt us." Science **328**(5980): 852-856.

- Sartain, M. J. and J. T. Belisle (2009). "N-Terminal clustering of the O-glycosylation sites in the Mycobacterium tuberculosis lipoprotein SodC." Glycobiology **19**(1): 38-51.
- Sartain, M. J., R. A. Slayden, K. K. Singh, S. Laal and J. T. Belisle (2006). "Disease state differentiation and identification of tuberculosis biomarkers via native antigen array profiling." Mol Cell Proteomics **5**(11): 2102-2113.
- Schubert, O. T., J. Mouritsen, C. Ludwig, H. L. Rost, G. Rosenberger, P. K. Arthur, M. Claassen, D. S. Campbell, Z. Sun, T. Farrah, M. Gengenbacher, A. Maiolica, S. H. Kaufmann, R. L. Moritz and R. Aebersold (2013). "The Mtb proteome library: a resource of assays to quantify the complete proteome of Mycobacterium tuberculosis." Cell Host Microbe **13**(5): 602-612.
- Scigelova, M. and A. Makarov (2006). "Orbitrap mass analyzer--overview and applications in proteomics." Proteomics **6 Suppl 2**: 16-21.
- Scott, N. E., B. L. Parker, A. M. Connolly, J. Paulech, A. V. Edwards, B. Crossett, L. Falconer, D. Kolarich, S. P. Djordjevic, P. Hojrup, N. H. Packer, M. R. Larsen and S. J. Cordwell (2011). "Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of Campylobacter jejuni." Mol Cell Proteomics **10**(2): M000031-mcp000201.
- Second, T. P., V. Zabrouskov, A. Makarov (2010). "Utilizing a hybrid mass spectrometer to enable fundamental protein characterization: intact mass analysis and top-down fragmentation with the LTQ Orbitrap MS". Thermo application note **498**
- Schlesinger, L. S., S. R. Hull and T.M. Kaufman. "Binding of the terminal mannosyl units of lipoarabinomannan from a virulent strain of Mycobacterium tuberculosis to human macrophages". J Immunol **152**: 2920-2930.
- Shen, Y., R. D. Smith, K. K. Unger, D. Kumar and D. Lubda (2005). "Ultrahigh-throughput proteomics using fast RPLC separations with ESI-MS/MS." Anal Chem **77**(20): 6692-6701.
- Shvartsburg, A. A., A. J. Creese, R. D. Smith and H. J. Cooper (2010). "Separation of peptide isomers with variant modified sites by high-resolution differential ion mobility spectrometry." Anal Chem **82**(19): 8327-8334.
- Sieling, P. A., P. J. Hill, K. M. Dobos, K. Brookman, A. M. Kuhlman, M. Fabri, S. R. Krutzik, T. H. Rea, D. G. Heaslip, J. T. Belisle and R. L. Modlin (2008). "Conserved mycobacterial lipoglycoproteins activate TLR2 but also require glycosylation for MHC class II-restricted T cell activation." J Immunol **180**(9): 5833-5842.
- Smith, G. T., M. J. Sweredoski and S. Hess (2014). "O-linked glycosylation sites profiling in Mycobacterium tuberculosis culture filtrate proteins." J Proteomics **97**: 296-306.
- Smith, L. M. and N. L. Kelleher (2013). "Proteoform: a single term describing protein complexity." Nat Meth **10**(3): 186-187.
- Spiro, R. G. (2002). "Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds." Glycobiology **12**(4): 43R-56R.

Stavrum, R., H. Valvatne, A. K. Stavrum, L. W. Riley, E. Ulvestad, I. Jonassen, T. M. Doherty and H. M. Grewal (2012). "Mycobacterium tuberculosis Mce1 protein complex initiates rapid induction of transcription of genes involved in substrate trafficking." Genes Immun **13**(6): 496-502.

Steentoft, C., S. Y. Vakhrushev, H. J. Joshi, Y. Kong, M. B. Vester-Christensen, K. T. Schjoldager, K. Lavrsen, S. Dabelsteen, N. B. Pedersen, L. Marcos-Silva, R. Gupta, E. P. Bennett, U. Mandel, S. Brunak, H. H. Wandall, S. B. Levery and H. Clausen (2013). "Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology." EMBO J **32**(10): 1478-1488.

Steentoft, C., S. Y. Vakhrushev, M. B. Vester-Christensen, K. T. Schjoldager, Y. Kong, E. P. Bennett, U. Mandel, H. Wandall, S. B. Levery and H. Clausen (2011). "Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines." Nat Methods **8**(11): 977-982.

Stover, C. K., P. Warrener, D. R. VanDevanter, D. R. Sherman, T. M. Arain, M. H. Langhorne, S. W. Anderson, J. A. Towell, Y. Yuan, D. N. McMurray, B. N. Kreiswirth, C. E. Barry and W. R. Baker (2000). "A small-molecule nitroimidazopyran drug candidate for the treatment of tuberculosis." Nature **405**(6789): 962-966.

Sutcliffe, I. C. and D. J. Harrington (2004). "Lipoproteins of Mycobacterium tuberculosis: an abundant and functionally diverse class of cell envelope components." FEMS Microbiol Rev **28**(5): 645-659.

Swaney, D. L., G. C. McAlister and J. J. Coon (2008). "Decision tree-driven tandem mass spectrometry for shotgun proteomics." Nat Methods **5**(11): 959-964.

Syka, J. E., J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt (2004). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." Proc Natl Acad Sci U S A **101**(26): 9528-9533.

Takegawa, Y., H. Ito, T. Keira, K. Deguchi, H. Nakagawa and S. Nishimura (2008). "Profiling of N- and O-glycopeptides of erythropoietin by capillary zwitterionic type of hydrophilic interaction chromatography/electrospray ionization mass spectrometry." J Sep Sci **31**(9): 1585-1593.

Thompson, A., J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann and C. Hamon (2003). "Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS." Analytical Chemistry **75**(8): 1895-1904.

Thomson, B. A. and J. V. Iribarne (1979). "Field induced ion evaporation from liquid surfaces at atmospheric pressure." The Journal of Chemical Physics **71**(11): 4451-4463.

Torrelles, J. B., A. K. Azad and L. S. Schlesinger (2006). "Fine discrimination in the recognition of individual species of phosphatidyl-myo-inositol mannosides from Mycobacterium tuberculosis by C-type lectin pattern recognition receptors." J Immunol **177**: 1805-1816.

Tortoli, E. (2006). "The new mycobacteria: an update." FEMS Immunol Med Microbiol **48**(2): 159-178.

Tran, J. C. and A. A. Doucette (2008). "Rapid and effective focusing in a carrier ampholyte solution isoelectric focusing system: a proteome prefractionation tool." J Proteome Res **7**(4): 1761-1766.

Tsur, D., S. Tanner, E. Zandi, V. Bafna and P. A. Pevzner (2005). "Identification of post-translational modifications by blind search of mass spectra." Nat Biotechnol **23**(12): 1562-1567.

Uetrecht, C., R. J. Rose, E. van Duijn, K. Lorenzen and A. J. Heck (2010). "Ion mobility mass spectrometry of proteins and protein assemblies." Chem Soc Rev **39**(5): 1633-1655.

Vissa, V. D. and P. J. Brennan (2001). "The genome of Mycobacterium leprae: a minimal mycobacterial gene set." Genome Biol **2**(8): REVIEWS1023.

Wang, H., C.-H. Wong, A. Chin, A. Taguchi, A. Taylor, S. Hanash, S. Sekiya, H. Takahashi, M. Murase, S. Kajihara, S. Iwamoto and K. Tanaka (2011). "Integrated mass spectrometry-based analysis of plasma glycoproteins and their glycan modifications." Nat. Protocols **6**(3): 253-269.

Washburn, M. P., D. Wolters and J. R. Yates, 3rd (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nat Biotechnol **19**(3): 242-247.

Wehmeier, S., A. S. Varghese, S. S. Gurucha, B. Tissot, M. Panico, P. Hitchen, H. R. Morris, G. S. Besra, A. Dell and M. C. Smith (2009). "Glycosylation of the phosphate binding protein, PstS, in Streptomyces coelicolor by a pathway that resembles protein O-mannosylation in eukaryotes." Mol Microbiol **71**(2): 421-433.

Whitehouse, C. M., R. N. Dreyer, M. Yamashita and J. B. Fenn (1985). "Electrospray interface for liquid chromatographs and mass spectrometers." Anal Chem **57**(3): 675-679.

Wiker, H.G, M. Harboe (1992). "The antigen 85 complex: a major secretion product of Mycobacterium tuberculosis." Microbiol Rev **56**(4): 648-661.

Wilm, M. and M. Mann (1996). "Analytical properties of the nanoelectrospray ion source." Anal Chem **68**(1): 1-8.

Wollscheid, B., D. Bausch-Fluck, C. Henderson, R. O'Brien, M. Bibel, R. Schiess, R. Aebersold and J. D. Watts (2009). "Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins." Nat Biotechnol **27**(4): 378-386.

Wuhrer, M., A. R. de Boer and A. M. Deelder (2009). "Structural glycomics using hydrophilic interaction chromatography (HILIC) with mass spectrometry." Mass Spectrom Rev **28**(2): 192-206.

Wynne, C., N. J. Edwards and C. Fenselau (2010). "Phyloproteomic classification of unsequenced organisms by top-down identification of bacterial proteins using capLC-MS/MS on an Orbitrap." Proteomics **10**(20): 3631-3643.

Xu, Y., Z. Wu, L. Zhang, H. Lu, P. Yang, P. A. Webley and D. Zhao (2009). "Highly specific enrichment of glycopeptides using boronic acid-functionalized mesoporous silica." Anal Chem **81**(1): 503-508.

Yates, J. R., D. Cociorva, L. Liao and V. Zabrouskov (2005). "Performance of a Linear Ion Trap-Orbitrap Hybrid for Peptide Analysis." Analytical Chemistry **78**(2): 493-500.

Zauner, G., C. A. Koeleman, A. M. Deelder and M. Wührer (2010). "Protein glycosylation analysis by HILIC-LC-MS of Proteinase K-generated N- and O-glycopeptides." J Sep Sci **33**(6-7): 903-910.

Zhang, H., X. J. Li, D. B. Martin and R. Aebersold (2003). "Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry." Nat Biotechnol **21**(6): 660-666.

Zubarev, R., N. L. Kelleher and F. McLafferty (1998). "Electron capture dissociation of multiply charged protein cations. A nonergodic process." J Am Chem Soc **120**: 3265-3266.

Zuber, B., M. Chami, C. Houssin, J. Dubochet, G. Griffiths and M. Daffe (2008). "Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state." J Bacteriol **190**(16): 5672-5680.

Zumla, A., P. Nahid and S. T. Cole (2013). "Advances in the development of new tuberculosis drugs and treatment regimens." Nat Rev Drug Discov **12**(5): 388-40

Title : Development of new proteomic strategies for the analysis of mycobacterial glycoproteins

Abstract: I developed some mass spectrometry methods for the search and the analysis of O-mannosylated proteins of *Mycobacterium tuberculosis*. A preliminary study in the laboratory showed the impact of the proteins glycosylation on the virulence of the pathogen

The first axis of the project consisted in the development of an innovative « bottom-up » proteomic approach to specifically search glycoproteins of *Mycobacterium tuberculosis*. For that, a dedicated bioinformatics tool was developed in the lab. I used this tool on data obtained after analysis of several proteic samples. With the use of this bioinformatic program and application of some filters we optimized, we could identify with confidence more than twenty glycoproteins of the pathogen.

Moreover, to fully characterize the glycosylation of the newly identified modified proteins, we also developed some mass spectrometry methods for the analysis of the whole proteins. I worked on a model protein, the Fasciclin of *Mycobacterium smegmatis*, we identified the glycosylation of in the lab.

Finally, this work allows to start a biological research program to evaluate the possible role of each of these newly identified glycoproteins on the virulence of the pathogen. Besides, we also performed a comparative and quantitative proteomic study of the wild type versus mutant Mtb strains to identify some metabolic ways that could be affected by the silencing of the $\Delta Rv1002c$ gene and to analyse the systemic impact of the protein-O-mannosylation deficiency on the pathogenicity of Mtb

Auteur : Laure TONINI

Titre : Développement de nouvelles stratégies protéomiques pour l'analyse de glycoprotéines mycobactériennes

Directeurs de thèse : Dr Odile SCHILTZ et Dr Michel RIVIERE, IPBS UMR 5089

Lieu et date de soutenance : 2 décembre 2014, amphithéâtre Gallais, LCC de Toulouse

Résumé : Je me suis intéressée au développement de méthodes d'analyses en spectrométrie de masse pour la recherche et la caractérisation des protéines *O*-mannosylées de *Mycobacterium tuberculosis* (Mtb). Une étude préliminaire faite au sein du laboratoire suggérait le caractère indispensable du phénomène de glycosylation pour la virulence du bacille. Dans un premier axe, j'ai développé une stratégie protéomique permettant la recherche de protéines glycosylées de Mtb au sein de mélanges complexes grâce à un programme informatique dédié développé au sein du laboratoire. Nous avons pu valider une liste d'une vingtaine de glycoprotéines de Mtb. Par ailleurs, afin de pouvoir caractériser de façon précise les profils et sites de glycosylation des glycoprotéines nouvellement identifiées de Mtb, j'ai également mis en place des méthodes d'analyse de glycoprotéines par spectrométrie de masse, en travaillant avec une protéine modèle de *Mycobacterium smegmatis*, la Fasciline. Différentes techniques, de l'analyse de la protéine entière à l'analyse des peptides m'ont permis de caractériser le profil de glycosylation de cette protéine. Enfin, dans le but de comprendre le rôle de ce processus dans la virulence de Mtb, j'ai cherché à identifier et caractériser les voies métaboliques affectées par l'arrêt de la glycosylation et potentiellement responsables de la perte de virulence chez le mutant $\Delta Rv1002c$, via une étude de protéomique quantitative. L'ensemble de ces travaux devrait permettre de contribuer de façon significative à la compréhension du rôle fonctionnel de la mannosylation des protéines dans la modulation des activités des mannoprotéines potentiellement impliquées dans la pathogénicité du bacille de la tuberculose.

Mots-clés : Protéomique - spectrométrie de masse - glycoprotéines - *O*-mannoprotéines - tuberculose - pathogénicité

Discipline : Biologie structurale et fonctionnelle

Laboratoire : Institut de Pharmacologie et de Biologie structurale, UMR 5089

